Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 45217, Pages 1–22
DOI 10.1155/ASP/2006/45217

An Automated Video Object Extraction System Based on Spatiotemporal Independent Component Analysis and Multiscale Segmentation

Xiao-Ping Zhang and Zhenhe Chen

Department of Electrical and Computer Engineering, Faculty of Engineering and Applied Science, Ryerson University, 350 Victoria Street, Toronto, ON, Canada M5B 2K3

Received 12 September 2004; Revised 13 March 2005; Accepted 27 May 2005

Video content analysis is essential for efficient and intelligent utilizations of vast multimedia databases over the Internet. In video sequences, object-based extraction techniques are important for content-based video processing in many applications. In this paper, a novel technique is developed to extract objects from video sequences based on spatiotemporal independent component analysis (stICA) and multiscale analysis. The stICA is used to extract the preliminary source images containing moving objects in video sequences. The source image data obtained after stICA analysis are further processed using wavelet-based multiscale image segmentation and region detection techniques to improve the accuracy of the extracted object. An automated video object extraction system is developed based on these new techniques. Preliminary results demonstrate great potential for the new stICA and multiscale-segmentation-based object extraction system in content-based video processing applications.

1. INTRODUCTION

The increasing popularity of video processing is due to the high demand for video in entertainment, security related applications, education, telemedicine, database, and new wireless telecommunications. Recently, interesting research topics such as automated and efficient content-based video processing techniques are attracting much attention. The content-based video presentation is an essential need for emerging broadcasting services, Internet, and security applications.

Raw video clips are usually binary streams that are not well organized. To represent their contents, video clips must be decomposed into objects so analysis can be performed. The object-based technique is one way of analyzing the video clips and it is gaining importance in achieving compression and performing content-based video retrieval.

Recently, partitioning video sequences into semantic video objects has been an active research area. Applications to object-based video representation include video conference, biomedical, surveillance, and content-based video indexing and retrieval. Video coding standard MPEG-4 also introduces an object-based framework for multimedia representation [1]. To maximize the benefit of the industry standard and to provide object-level multimedia interaction, automatic video object segmentation techniques need to be developed.

Classical solutions to video object segmentation are based on motion features. A technique to represent video in layers was proposed in [2]. Image sequence is decomposed into layers by estimating and clustering affine parameters. Borshukov et al.[3] improved this method by replacing adaptive K-means with a merging algorithm and implementing the block-based affine modeling in multistage. A modified Hough transform [4] and a Bayesian framework [5] were also proposed in the literature for motion segmentation.

Spatiotemporal information could be used for video object segmentation. In [6], a region-merging approach is proposed to identify video objects. This method starts from an oversegmentation of the current frame and then iteratively merges the regions based on spatiotemporal similarity. Temporal similarity is estimated by a modified Kolmogorov-Smirnov test. In [7], an algorithm based on higher-order statistics significance test was described to separate moving video objects from background. Kim and Hwang [8] utilized edge change information to extract video objects. Another spatiotemporal segmentation approach based on edge flow and 3D motion estimation was proposed in [9]. Other techniques that combine video object segmentation and tracking were proposed in [10–12]. Performance could be improved

by integrating multiple features [13, 14], user interaction [15–17], and multiview extensive partition operators [18]. Due to the limitation of motion estimation, motion segmentation techniques may not give accurate object boundaries. For nonrigid objects, active contour (i.e., snakes) models have been widely used for image segmentation. However, in order to successfully solve the active contour models, it is very important to have accurate initializations [15].

Spatiotemporal segmentation techniques consider both spatial and temporal information. For top-down spatiotemporal segmentation algorithms, motion parameters may not be accurately estimated due to imperfect outlier detection. The bottom-up spatiotemporal segmentation techniques typically consist of a spatial segmentation step and a merging step based on temporal information. Even though both spatial and temporal information are considered during processing, spatial information and temporal information are used in separate stages. Also, most algorithms only utilize two successive frames.

In recent years, the independent component analysis-(ICA-) based techniques are getting much attention in video processing. The ICA can be used in two complementary ways to decompose an image sequence into a set of images and a corresponding set of time-varying image amplitudes. The spatial ICA (sICA) [19] finds a set of mutually independent component (IC) images and a corresponding set of unconstrained time courses, whereas the temporal ICA (tICA) [20] finds a set of IC time courses and a corresponding set of unconstrained images. However, the sICA and tICA can only seek either the ICs of images (frames) or the time courses, respectively. As shown in [19], the sICA extracts the independent images but the corresponding temporal sources could be highly correlated, while tICA only extracts independent temporal sources but not independent images. This is undesirable for object-based video sequence analysis, since the corresponding time courses for the independent objects should be independent as well. The stICA, the generalization of the classic ICA, was initially developed in functional magnetic resonance imaging (fMRI) [21]. It can blindly separate the independent sources from their spatial and temporal mixtures.

In this paper, a systematic framework is presented for automated content-based video processing based on the spatiotemporal independent component analysis (stICA) and multiscale analysis. First, a novel stICA model is used to formulate the spatial and temporal independence of various moving objects. The solution of the stICA model can therefore identify these objects. In the new algorithm, areas of video objects are extracted without explicitly performing spatial and motion segmentation. The new algorithm takes multiple frames as input, and then finds the spatial and temporal independence simultaneously. Multiple moving objects are extracted at the same time. The independent component with highest energy is considered to be the background. Postprocessing based on multiscale region segmentation and other analysis is also introduced to refine video object boundaries. A new iterative algorithm is also presented to solve the nonlinear combination problem of the stICA modeling of video sequences. Both theoretical derivation and simulation results are given to illustrate the effectiveness of the presented methods.

The main contributions of this paper include: (i) a new method to analyze video sequences based on the stICA model; (ii) a novel compensation method to deal with the nonlinear combination problem in the stICA model for video sequences; (iii) the integrated postprocessing techniques based on wavelet analysis, edge detection with region growing, and multiscale segmentation approaches.

The paper is organized as follows. Section 2 introduces the framework of the proposed new automated video object extraction system based on a new formulation of the stICA model for video object extraction. Section 3 describes the algorithms of the first iteration of the stICA-based video segmentation, including the postprocessing based on multiscale region segmentation. In Section 4, a new compensation approach is presented to solve the nonlinear combination problem for the practical video stICA model, which is the basis of the second iteration of the stICA-based video segmentation. Extensive simulation results are presented in Section 5 to illustrate the effectiveness of the algorithms in each component of the system. Finally, Section 6 concludes the paper.

2. FRAMEWORK OF A NEW AUTOMATED VIDEO OBJECT EXTRACTION SYSTEM USING ICA AND MULTISCALE ANALYSIS

2.1. An stICA model for video object extraction

2.1.1. Independent component analysis (ICA) and spatiotemporal independent component analysis (stICA)

ICA is a statistical technique introduced in the 1980s [22] in the context of neural network modeling. The purpose of ICA is to restore statistical independent source signals given only observed output signals without knowing the mixing matrix or sources. Comparing to principle component analysis (PCA) [23] which solves the correlation problems, ICA can reduce the high-order dependencies to make the sources as independent as possible. ICA technique is based on a mixing model given by

$$\mathbf{X} = \mathbf{AS}, \tag{1}$$

where there are M observations and the $M \times N$ output observation matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, with $\mathbf{x}_i$, $i = 1, \ldots, N$, being $M \times 1$ column vectors and N the number of samples, and unknown source matrix $\mathbf{S} = [\mathbf{s}_1^T, \mathbf{s}_2^T, \ldots, \mathbf{s}_K^T]^T$, where $N \times 1$ unknown column vectors $\mathbf{s}_i$, $i = 1, \ldots, K$, are K independent unknown source vectors. The matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_K]$ is an $M \times K$ unknown mixing matrix, where $M \times 1$ column vectors $\mathbf{a}_i$, $i = 1, \ldots, K$, are the mixing signature for source $\mathbf{s}_i$.

There are two constraints in the ICA model: (i) source signals $\mathbf{s}$ must be non-Gaussian, and (ii) the components of $\mathbf{s}$ are statistically independent. If the mixture signals can be decomposed into non-Gaussian and statistically independent signals, these independent signals form the estimation of source signals.

If the observed samples are temporal samples, that is, s_{in}, $n = 1,\ldots,N$, are temporal sample sequences from time 1 to N for the independent spatial source i, $i = 1,\ldots,K$, formulation (1) becomes the spatial ICA (sICA).

Taking a transpose of (1), denoted by the superscript "T," we have

$$\mathbf{X}^T = \mathbf{S}^T\mathbf{A}^T. \tag{2}$$

Now, $\mathbf{S}$ also looks like a mixing matrix. If the columns of matrix $\mathbf{A}^T$ are assumed statistically independent, a temporal ICA (tICA) problem is formulated since the row vectors of $\mathbf{A}^T$ correspond to the columns of $\mathbf{S}^T$, representing the time courses of the signal source. Note that in the tICA, the independence of spatial sources in $\mathbf{S}$ is not assumed.

Apparently, the sICA and tICA only seek either ICs of images (frames) or time courses, respectively [19]. The sICA extracts independent images but the corresponding temporal sources could be highly correlated, while the tICA only extracts independent temporal sources but not independent images.

However, for object-based video sequences analysis, both objects and the corresponding time courses for the objects can be assumed independent, that is, both the row vectors of $\mathbf{S}$ and the row vectors of $\mathbf{A}$ are independent. Therefore, an stICA model may be formulated. In stICA, not only spatial source signals (images) are a set of ICs, but the time courses should also be a set of ICs. The stICA, the generalization of classic ICA, can blindly separate the independent sources from their spatial and temporal mixtures. It was initially developed in functional magnetic resonance imaging (fMRI) [21]. For clarity and simplicity, the stICA is formulated as follows (note that the notations are different from (1)).

Let $M \times N$ matrix contain a sequence of n images $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_N]$. Each image $\mathbf{x}_i$ is an $M \times 1$ vector. A linear decomposition of $\mathbf{X}$ can be represented by a matrix factorization,

$$\mathbf{X} = \mathbf{S}\Lambda\mathbf{T}^T, \tag{3}$$

where the $M \times K$ matrix $\mathbf{S} = [\mathbf{s}_1,\ldots,\mathbf{s}_K]$ represents the spatial image source sequence and the $M \times 1$ column vectors $\mathbf{s}_i$, $i = 1,\ldots,K$, represent unknown independent image sources. In the mixing matrix $\mathbf{T} = [\mathbf{t}_1,\ldots,\mathbf{t}_K]$, the $N \times 1$ column vectors $\mathbf{t}_i$ represent the corresponding independent time courses for different sources, that is, it is assumed that difference image sources have unknown independent time courses. The matrix Λ is a diagonal matrix of scaling parameters. Note that in ICA problems, Λ is irresolvable without other prior information [24]. To solve both $\mathbf{S}$ and $\mathbf{T}$ when only the mixture observation $\mathbf{X}$ is known, the following procedures are employed.

Singular value decomposition (SVD) [23] can reduce the rank of mixture and factorize it as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{4}$$

where $\mathbf{U}$ is an $M \times K$ matrix of $K \le M$ eigenimages, $\mathbf{V}$ is a $N \times K$ matrix of $K \le N$ eigensequences, and $\mathbf{D}$ is a diagonal matrix of singular values. In order to determine the independent $\mathbf{S}$ and $\mathbf{T}$, two $K \times K$ unmixing matrices $\mathbf{W}_S$ and $\mathbf{W}_T$ are assumed to exist such that

$$\mathbf{S} = \tilde{\mathbf{U}}\mathbf{W}_S, \tag{5}$$

$$\mathbf{T} = \tilde{\mathbf{V}}\mathbf{W}_T, \tag{6}$$

where $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{D}^{1/2}$ and $\tilde{\mathbf{V}} = \mathbf{V}\mathbf{D}^{1/2}$. Now we have

$$\mathbf{X} = \mathbf{S}\Lambda\mathbf{T}^T = \tilde{\mathbf{U}}\mathbf{W}_S\left(\tilde{\mathbf{V}}\mathbf{W}_T\right)^T = \tilde{\mathbf{U}}\mathbf{W}_S\mathbf{W}_T{}^T\tilde{\mathbf{V}}^T. \tag{7}$$

To find the unmixing matrices $\mathbf{W}_T$ and $\mathbf{W}_S$, the following informax principle is applied [20, 25]. The independent spatial and temporal components are expected to simultaneously maximize a function h_{ST} of the spatial entropy

$$h_S = H(\sigma(\tilde{\mathbf{U}}\mathbf{W}_S)), \tag{8}$$

and temporal entropy

$$h_T = H(\tau(\tilde{\mathbf{V}}\mathbf{W}_T)), \tag{9}$$

where σ and τ approximate the cumulative density function (cdf) of each of the spatial source signals and temporal signals, respectively. The function h to be maximized is defined as

$$h_{ST}(\mathbf{W}_S) = \alpha h_S + (1-\alpha)h_T, \tag{10}$$

where α is a weighting factor given to spatial and temporal entropy. To optimize these two entropies by maximum likelihood estimation [20], their notations need to be changed to

$$\begin{aligned} h_S &= \log|\mathbf{W}_S| + \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{k}\log\sigma_i'(s_{ij}), \\ h_T &= \log|\mathbf{W}_T| + \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{k}\log\tau_i'(t_{ij}), \end{aligned} \tag{11}$$

where s_{ij} and t_{ij} are the corresponding elements of $\mathbf{S}$ and $\mathbf{T}$ in (3). σ_i and τ_i are the cdfs of the spatial and temporal signals, respectively. Their derivatives σ_i' and τ_i' are the corresponding pdfs.

One can recover the spatial signals and the time courses at the same time using maximum likelihood estimation, which is similar to the conventional ICA [26] approximation techniques.

2.1.2. *Formulation of the stICA model for video sequences*

Let us denote a video sequence with N frames as $\hat{\mathbf{F}} = [\hat{\mathbf{f}}_1,\ldots,\hat{\mathbf{f}}_N]$, where $\hat{\mathbf{f}}_i$ is an $M \times 1$ column vector representing a frame that contains M pixels. These image vectors are constructed by taking the column-wise elements from the frame images. Thus the dimension of matrix $\hat{\mathbf{F}}$ is $M \times N$. The mutual independent objects of interest are denoted as $\mathbf{O} = [\mathbf{o}_1,\ldots,\mathbf{o}_K]$, where $\mathbf{o}_i$ is constructed in the same way as

$\hat{\mathbf{f}}_i$ and $K \leq N$. The dimension of the object vector $\mathbf{o}_i$ is $M \times 1$, the same as $\hat{\mathbf{f}}_i$. Thus, the dimension of $\mathbf{O}$ is $M \times K$. If the video sequence is captured by a fixed camera, such as in the surveillance security system, the background is a constant. The stationary background can be considered as a vector of $\mathbf{O}$, say $\mathbf{o}_K$. The independent temporal signals (time courses) $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$ affect the objects on every time unit. Again, we use the same method to construct the time course column vector $\mathbf{a}_i$. In every time unit, there are time courses affecting each object and the dimension of any time course vector $\mathbf{a}_i$ should be equal to the number of video frames, that is, $N \times 1$. This means that each column of $\mathbf{A}$ is the time signature for the corresponding objects in $\mathbf{O}$. The dimension of $\mathbf{A}$ is $N \times K$, where $K \leq N$. Because the background is stationary, the corresponding time course vector $\mathbf{a}_K$ has no effect on it and all elements of vector $\mathbf{a}_K$ have value 1. We have

$$\hat{\mathbf{F}} = \mathbf{O}\mathbf{A}^T. \tag{12}$$

Note that given the spatial and temporal independence assumptions, (12) exactly fits into the stICA model in (3), where the independent spatial source matrix $\mathbf{S}$ in (3) is replaced by the independent spatial object images $\mathbf{O}$ in (12) and the independent time courses $\mathbf{T}$ by the independent time courses $\mathbf{A}$.

To find out the effect of each object on the video frames, we expand the matrices

$$\begin{aligned}
&[\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots, \hat{\mathbf{f}}_N] \\
&\quad = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_K][\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_K]^T \\
&\quad = \begin{bmatrix} o_{11} & o_{12} & \cdots & o_{1K} \\ \vdots & \vdots & \cdots & \vdots \\ o_{M1} & o_{M2} & \cdots & o_{MK} \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{N1} \\ \vdots & \vdots & \cdots & \vdots \\ a_{1K} & a_{2K} & \cdots & a_{NK} \end{bmatrix} \\
&\quad = \begin{bmatrix} a_{11}o_{11} & a_{21}o_{11} & \cdots & a_{N1}o_{11} \\ \vdots & \vdots & \cdots & \vdots \\ a_{11}o_{M1} & a_{21}o_{M1} & \cdots & a_{N1}o_{M1} \end{bmatrix} \\
&\qquad + \begin{bmatrix} a_{12}o_{12} & a_{22}o_{12} & \cdots & a_{N2}o_{12} \\ \vdots & \vdots & \cdots & \vdots \\ a_{12}o_{M2} & a_{22}o_{M2} & \cdots & a_{N2}o_{M2} \end{bmatrix} \\
&\qquad + \cdots \\
&\qquad + \begin{bmatrix} a_{1K}o_{1K} & a_{2K}o_{1K} & \cdots & a_{NK}o_{1K} \\ \vdots & \vdots & \cdots & \vdots \\ a_{1K}o_{MK} & a_{2K}o_{MK} & \cdots & a_{NK}o_{MK} \end{bmatrix}.
\end{aligned} \tag{13}$$

A function g is assumed to describe the object $\mathbf{o}_i$'s contribution to $\hat{\mathbf{F}}$. From the above matrices expansion, we can see that

$$g(\mathbf{o}_i) = \mathbf{o}_i\mathbf{a}_i^T, \quad i = 1, \ldots, K. \tag{14}$$

These equations reveal the fact that $\mathbf{a}_i$ is the time signature for the corresponding object $\mathbf{o}_i$. We can rewrite (12) in vector format as

$$\hat{\mathbf{F}} = \sum_{i=1}^{K} g(\mathbf{o}_i) = \sum_{i=1}^{K} \mathbf{o}_i\mathbf{a}_i^T. \tag{15}$$

To find the element construction in jth video frame $\hat{\mathbf{f}}_j$, $j = 1, \ldots, K$, we need to utilize the linear combination relationship between the spatial elements o_{ik} and the time sequence signals a_{jk} from (12) and (13):

$$\hat{f}_{ij} = \sum_{k=1}^{K} o_{ik}a_{jk}, \tag{16}$$

where $i = 1, \ldots, M$. This equation shows that an element at a specific location in a frame is the linear combination of the elements at the same locations of all the independent spatial objects at a certain time moment i, that is, the ith element in the jth video frame is the linear combination of all the ith elements in all the independent object vectors $\mathbf{o}_1, \ldots, \mathbf{o}_K$ at ith moment.

Figure 1 demonstrates how the stICA model is applied to video frames. At a certain moment, a video frame consists of a linear combination of all objects, including the background. For example at $t = 1$, video frame 1 is obtained by the linear combination of the spatial ICs on the left-hand side of Figure 1. Frame 1 carries the information of the background, object 1 and object 3. In this way, different video frames are constructed.

Note that the video frames actually are not the linear combinations of the ICs as we wish because moving objects block (not add on) the background in the video frames. This condition violates the stICA assumption. Thus, we need to compensate for the background information that is lost due to object blocking. In this way, the assumption of linear combination may hold so that the stICA requirements are satisfied. Here we denote the ideally blocked background information by Δ_i in ith frame $\mathbf{f}_i$, such that

$$\hat{\mathbf{f}}_i = \mathbf{f}_i + \Delta_i, \tag{17}$$

where the dimension of Δ_i is also $M \times 1$ and $i = 1, \ldots, N$.

Between the practical video frame model in (17) and the fitting model in (15), there is a gap Δ_i that affects the accuracy of the stICA approach on video sequences. This problem is dealt with by a novel compensation method presented in Section 4.

2.2. A new generic video object segmentation system based on stICA and multiscale analysis

Based on the stICA model formulated in the above section, a new generic video object segmentation system is developed. Figure 2 shows the main algorithmic modules of this system in the block diagram.

The main algorithm includes two iterations. Both iterations employ the stICA model and associated algorithm for

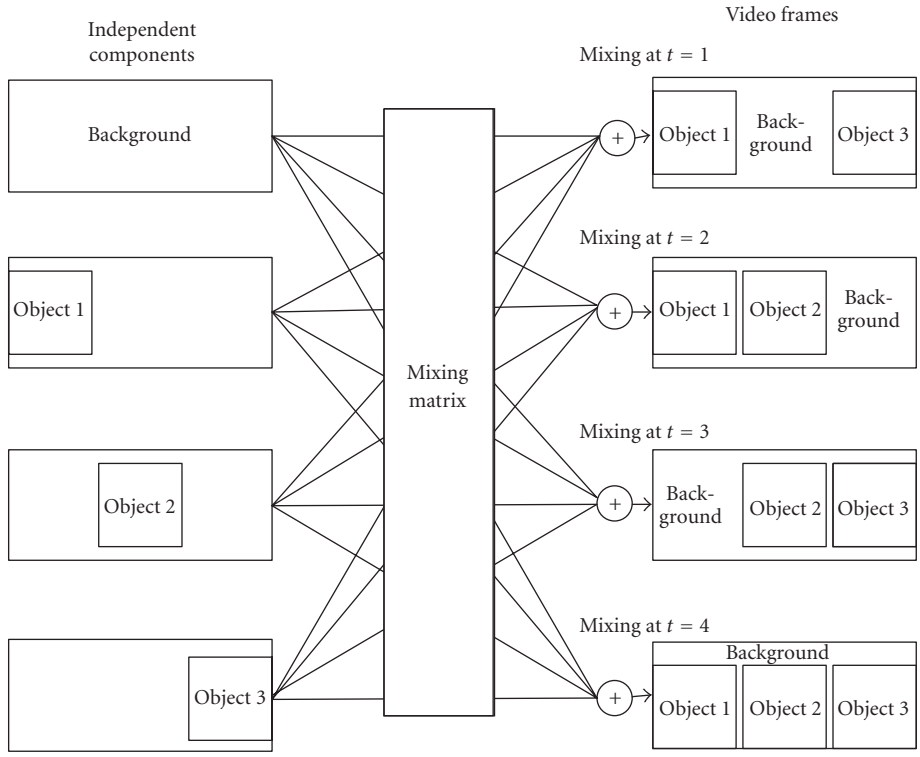


Figure 1: Illustration of video frame construction by mixing objects.

background and object extraction. Similar post-processing procedures are also employed in both iterations. Two iterations are summarized as follows.

First iteration

The stICA model and associated algorithm are applied to video frames to separate the spatial and temporal signals. In our system, video frames are selected as observed mixture signals $\mathbf{x}$, and an informax algorithm [20, 25] is applied on these signals to extract the preliminary signals that represent objects.

The signals obtained after the stICA are further processed. The wavelet analysis, edge detection with region growing, and multiscale image segmentation techniques are employed to refine the extracted preliminary objects of interest.

Second iteration

A compensation approach is introduced to deal with the nonlinear combination problem of the stICA. The blocked background is compensated based on the object extraction results in iteration 1 (see (17)). The procedures of stICA and post-processing in iteration are reapplied on the compensated observations. A frame object indexing technique is then performed to reconstruct the sequence of frames containing only the objects. More precise video objects are extracted in this iteration.

Each algorithmic module is described in the following sections.

3. THE stICA-BASED VIDEO SEGMENTATION: THE FIRST ITERATION

The first iteration includes the following steps (as shown in Figure 3).

(1) Use the stICA to process selected frames from a video sequence. The preliminarily processed images are obtained by subtracting the recovered background from original video frames.

(2) The preliminarily processed images are processed by using the wavelet-analysis-based nonlinear detector to obtain the rectangular regions of interest (ROIs).

(3) From the ROIs, edge detection of the extracted object is performed. A recursive region growing technique is employed to remove the small-size regions in the ROIs. The object regions are formed in this step.

(4) Multiscale segmentation techniques are applied to the object regions with the eroding/projecting approach to identify the regions belonging to the object.

3.1. Initial object segmentation based on stICA model

In the first iteration (block diagram in Figure 3), the stICA model is applied to the captured video frames. According to the stICA model described in (12), (13), (14), (15), and (16), the video sequences $\hat{\mathbf{F}}$ are used directly as the observed image

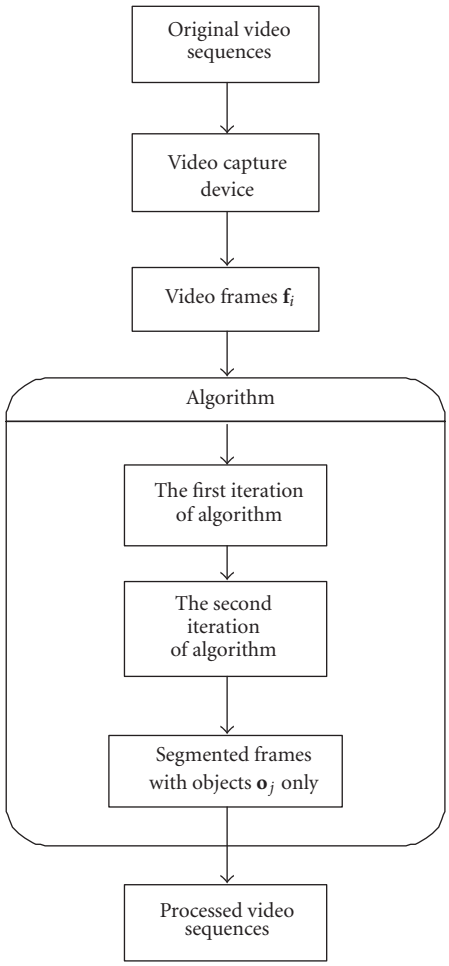


Figure 2: Block diagram of the system framework, where i and j are the indices of frames and objects, respectively.

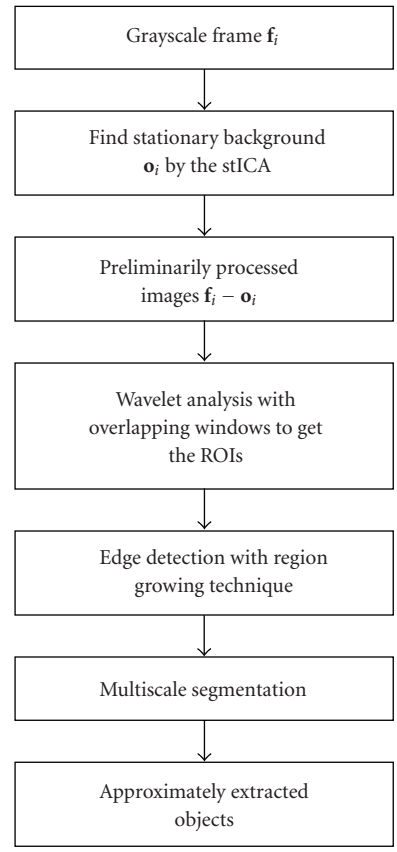


Figure 3: Block diagram of the first iteration.

matrix $\mathbf{X}$, to which the stICA algorithm is applied. According to (3), (4), (5), (6), and (7), video sequence frames can be decomposed into two parts by SVD, eigenimages and corresponding eigen-time courses:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \left(\mathbf{U}\mathbf{D}^{1/2}\right)\left(\mathbf{V}^T\mathbf{D}^{1/2}\right) = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^T. \tag{18}$$

The objective is to find the unmixing matrices $\mathbf{W}_S$ and $\mathbf{W}_T$ and the ICs, $\mathbf{S}$ and $\mathbf{T}$.

The informax criterion [20, 25] represented by (8), (9), (10), and (11) is employed on both spatial and temporal matrices. Conjugate gradient minimization [27] is implemented to find the unmixing matrices $\mathbf{W}_S$ and $\mathbf{W}_T$ and the ICs, $\mathbf{S}$ and $\mathbf{T}$. The maximum likelihood estimation is employed on both spatial and temporal signals. The informax-based algorithm [20, 25] is implemented to find the unmixing matrices in (3), (4), (5), (6), and (7).

However, since the video frames are not linear combinations of objects and the objects are not exactly ICs, the recovered spatial signals $\mathbf{o}_i$ are still coarse representation of the objects. The stICA approach alone cannot provide a satisfactory object segmentation result. Postprocessing techniques are then required to refine the object segmentation. The postprocessing procedures in the first iteration are illustrated in Figure 3. The inputs for postprocessing are the preliminarily processed images obtained by subtracting the recovered background by stICA from the original video frames. The wavelet analysis is employed to locate the rectangular ROIs. Then the edge detection and region growing approaches are used to extract object boundaries and region edges. Subsequently, the small-size regions are isolated and removed. After the edge detection and the region growing, there may still be some superfluous connected components with similar grayscale to the real objects. To remove the superfluous connected components from an object, the multiscale segmentation technique is applied to the object regions. Through the presented eroding and projecting approaches, multiscale segmented regions belonging to the object can be identified.

In the following subsections, we present these postprocessing techniques sequentially.

3.2. Using wavelet analysis to locate ROIs

As a powerful tool of image analysis, the wavelet transform performs well in characterizing singularities [28, 29]. In other words, large coefficients represent edge transitions in the wavelet domain. As known, the 2D discrete wavelet transform (DWT) decomposes an image into three wavelet

subspaces, namely, LH, HL, and HH, and one scaling subspace LL, where letter "L" means lowpass filter in the DWT and letter "H" means the bandpass filter in the DWT. The first letter represents horizontal direction and the second letter represents vertical direction.

The HL subspace is used to detect the horizontal boundaries of image objects and the LH subspace to detect the vertical boundaries. In the HL subspace, horizontal edges are represented by large coefficients. A horizontal sliding window filter is applied to detect the coefficient with the largest absolute value which may represent the horizontal boundary of the object. Thus the horizontal scope of image objects can be detected in the HL subspace and the horizontal region of interest (denoted by $\mathrm{ROI}_{\mathrm{HL}}^{\mathrm{horizontal}}$) may be identified in the wavelet domain.

For any spatial signal after the stICA processing, we denote $\mathbf{W}$ as the HL subspace at the Nth level of the wavelet decomposition and w_{ij} is the coefficient in that subspace, where i, j are the indices of rows and columns of $\mathbf{W}$, respectively. The following major procedures are involved in the algorithm.

Step 1. A row vector $\mathbf{\Psi}[\psi_1, \ldots, \psi_q]$ is obtained to represent the ensemble of those largest coefficient values in the columns of the HL subspace, that is, $\psi_j = \max_i |w_{ij}|$. An example of this vector is shown in Figure 4(a). Note that q is the total number of columns in the subspace, determined by the level of the wavelet decomposition. For example, if the dimension of an image is $r \times r$, then

$$q = \left(\frac{1}{2}\right)^N \times r. \tag{19}$$

Step 2. An overlapping sliding window filter with width l is used. The mean value of the largest absolute values ψ_i within the window is calculated:

$$m_k = \frac{\sum_{i=k}^{k+l-1} \psi_i}{l}, \quad k = 1, \ldots, q - l + 1. \tag{20}$$

An example filtering result is shown in Figure 4(b). A threshold filtering is then employed to segment the object area from the background:

$$m'_k = \begin{cases} m_k, & m_k \geq \max\{\psi_1, \ldots, \psi_q\} \times \alpha \\ & \quad = \max_i \{\max_j |w_{ij}|\} \times \alpha, \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

where α is an empirical constant, $k = 1, \ldots, q + l - 1$, and $i, j = 1, \ldots, r$. An example of the thresholding result is shown in Figure 4(c). Denote the maximum horizontal range of continuous nonzero m'_k as $[a, b]$. The horizontal ROI is detected as:

$$\mathrm{ROI}_{\mathrm{HL}}^{\mathrm{horizontal}} = \{i \mid a \leq i \leq b\}, \tag{22}$$

where i is the column index. In this way, all (may be multiple) regions containing object edge can be detected.

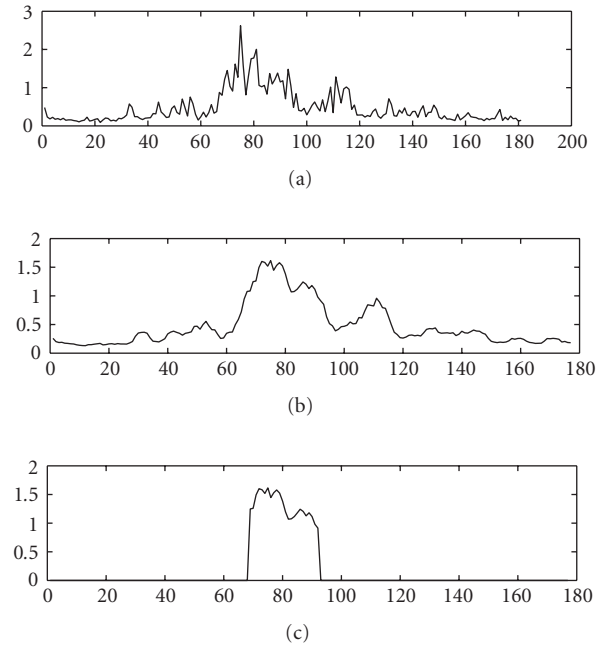


Figure 4: (From top to bottom): (a) maxima of absolute values of wavelet coefficients; (b) mean values of the maxima in the overlapping sliding windows; (c) mean values after threshold filtering.

The same algorithm is applied in the LH subspace, the vertical ROI can then be detected:

$$\mathrm{ROI}_{\mathrm{LH}}^{\mathrm{vertical}} = \{j \mid c \leq j \leq d\}, \tag{23}$$

where j is the row index, and c, d are the starting and ending points of vertical edges, respectively. Thus, the rectangular, ROIs that contain the objects in the wavelet domain are obtained as

$$\mathrm{ROI}_{\mathrm{wavelet}} = \left\{i, j \mid i \in \mathrm{ROI}_{\mathrm{HL}}^{\mathrm{horizontal}},\ j \in \mathrm{ROI}_{\mathrm{LH}}^{\mathrm{vertical}}\right\}. \tag{24}$$

The corresponding ROI in the stICA processed images can be located by using the inverse calculation in (19).

The purpose of segmenting an ROI is to decrease computational complexity for later postprocessing and to reduce noise so that edge detection techniques and region-based segmentation approaches can achieve better results. After ROI detection, the edge detection with region growing combined with a multiscale image segmentation is employed to identify accurate objects within ROI.

3.3. Image edge detection with region growing

The ROIs detected by the presented object detection method based on the stICA represent areas of the objects of interest. However, they do not contain boundary information of the objects. The Canny edge detection technique [30] is then applied to these rectangular ROIs to detect the closed regions for possible object regions. A binary image is rendered by the Canny edge detection. The interior regions inside the closed edge are represented by the value 1.

Not all the obtained regions contain objects of interest. In the ROIs, the target objects are generally larger than other isolated regions. Thus we can discriminate the target objects from those unwanted regions through comparing their sizes. A region growing method based on the basic procedures in [31] is employed to calculate the connected region size, briefly summarized as follows.

In a binary image **I**, two pixels are considered to be in the same region if they are in their neighbors of eight and have same grayscale value.

Two matrices, "Mark Matrix" and "Label Matrix," are defined to implement the region growing. All pixel values in the two matrices are initialized to zero. The flag with value 1 is assigned to a certain pixel in the Mark Matrix **M** to indicate that this pixel has been processed to avoid repeated processing. The Label Matrix **L** is used to assign a unique labeling integer to each isolated region. Thus the isolated regions can be distinguished by the different labeling integers. The total number of each labeling integer indicates the region size. Figure 5 shows an example.

A region label needs to be assigned for each pixel in an ROI. First of all, in the binary image **I**, a seed pixel $I_{i,j}$ is selected, which must satisfy two criteria:

(1) pixel value must be 1: $I_{i,j}$=1;
(2) the Mark Matrix element value cannot be 1 : $M_{i,j} \neq 1$. Otherwise, $I_{i,j}$ has been processed.

Once a new seed pixel $I_{i,j}$ is chosen, its eight neighbors $I_{p,q}(|p-i| \leq 1,\ |q-j| \leq 1)$ are examined. There are two underlying possibilities.

(1) If $I_{p,q} = 1$, for all $p \neq i, q \neq j, |p-i| \leq 1, |q-j| \leq 1$, the value of the corresponding element in the Mark Matrix **M**, $M_{p,q}$ should be checked. There are two possibilities under this condition.
    (a) $M_{p,q} = 1$: this indicates that the pixels corresponding to $M_{p,q}$ and $I_{p,q}$ have been processed. Thus, $I_{i,j}$ belongs to the same region as $I_{p,q}$, and $L_{i,j}$ is assigned the same value as $L_{p,q}$.
    (b) $M_{p,q} = 0$: this implies that $I_{p,q}$ has not been processed. If all the eight neighbors of $I_{i,j}$ have not been processed, $L_{p,q}$ and $L_{i,j}$ are both assigned a new labeling integer.
(2) If $I_{i,j}$ is the only pixel with value 1 in its region, $M_{i,j}$ is flagged to 1 and $L_{i,j}$ is assigned a new labeling integer.

In this way, all $I_{i,j}$'s neighbors $I_{p,q}$ with value 1 are identified. Their Mark Matrix elements $M_{i,j}$, $M_{p,q}$ are marked flag 1 after they have been processed. The corresponding Label Matrix elements $L_{i,j}$ and $L_{p,q}$ are assigned the same labeling integer.

This recursive region growing method identifies all isolated regions with different labeling integers by the Label Matrix **L**. A region-size threshold detector is used to remove regions with small sizes, which are not the objects of interest.

In this postprocessing step, we apply the Canny edge detection technique to the rectangular ROIs and then exploit the region growing method to remove small regions that are not objects of interest. The approximate object regions with boundaries are identified.

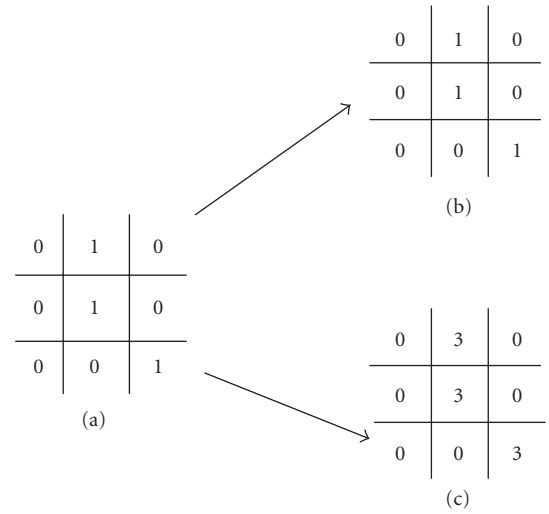


Figure 5: Region growing technique to label connected pixels. (a) Binary edge pixel neighborhood **I**; (b) mark pixel neighborhood **M**; (c) label pixel neighborhood **L**.

3.4. Multiscale image segmentation

Edge detection techniques such as the Canny method work efficiently on sharp edges. However, the processed images after the stICA usually do not possess sharp edges. This leads to some false edges that affect further processing.

In Figure 6, the objects of interest are obtained by edge detection with region growing to remove the small regions that are disconnected with the objects. However, this approach cannot remove the regions that are connected to the objects. For simplicity, these connected regions are called "connected components." Because of the false edges generated by edge detection, the region growing method cannot accurately identify the edges. Thus, a multiscale region-based still-image segmentation method [32–35] is employed on the object regions in postprocessing. Note that here the term "multiscale" means the scales of the grayscale variance in a region. A region in this method is a homogeneous region, which is defined as a connected region with a closed boundary and certain grayscale variance. Each region is labeled with a unique integer.

Apparently, segmentation of homogeneous regions with similar grayscale generally does not segment the objects of interest in images. A grayscale region may contain multiple objects, or one object may be divided to several grayscale regions. If an image has complex structure, it is difficult to find correspondence between each closed homogeneous region and a specific object. In Figure 6(c), an object and its connected component are divided to four homogeneous regions (R_1, R_2, R_3, and R_4) according to their grayscale similarities. In this case, homogeneous regions R_1, R_2, and R_3 belong to the object of interest. However, we cannot segment R_1, R_2, and R_3 from R_4 if using only multiscale segmentation.

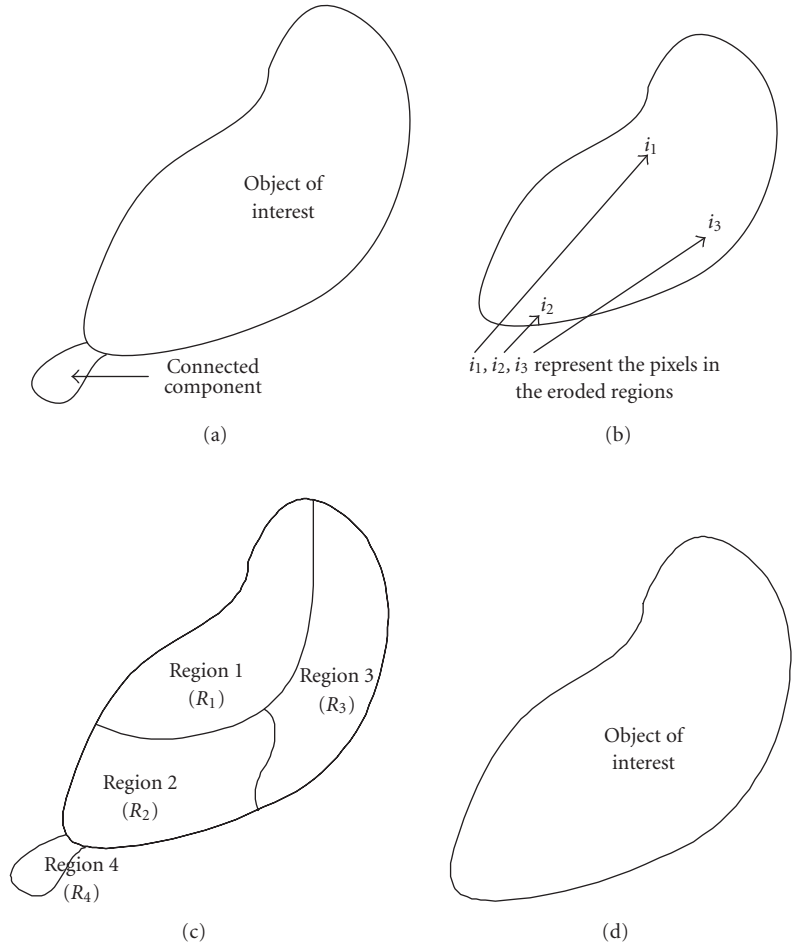


FIGURE 6: Illustration of the procedures of incorporating edge detection and multiscale segmentation: (a) regions obtained by edge detection and region growing; (b) eroded regions; (c) regions obtained by multiscale segmentation; (d) objects obtained by the projecting operation between (b) and (c).

We apply an eroding [30] and projecting approach on the multiscale segmentation results to obtain the objects of interest. The underlined (reasonable) assumption is that the connect components (which are not part of the object) are relatively small regions such that eroding will effectively remove them. The eroding results are shown in Figure 6(b). Afterwards, the eroded region R_e is combined with the multiscale segmentation results. The R_n is classified to be in the object area R_o, that is, $R_n \in R_o$ only if

$$\text{Area}\left(R_n \bigcap R_e\right) > 0.5\,\text{Area}\left(R_n\right), \tag{25}$$

where the operator $\text{Area}(\cdot)$ calculates the area of a region. After all R_n have been classified, final object areas are determined to be

$$R_o = \bigcup_n \left(R_n \in R_o\right). \tag{26}$$

In this way, the appropriate homogenous regions contained in the objects of interest are found with the exact boundaries identified, as illustrated in Figure 6(d).

In this way, by utilizing wavelet analysis, edge detection, region growing, and multiscale image segmentation approaches on the stICA outputs, objects with shape and boundaries can be extracted.

4. A COMPENSATION APPROACH OF stICA FOR PRACTICAL VIDEO SEQUENCES: THE SECOND ITERATION

When the background is complex enough, the linear stICA model may lead to inaccurate background identification in the first place (as described in Section 2.1.2), and therefore affect the subsequent processing. To deal with this problem and the nonlinear combination problem in the stICA model for video sequences, a novel "compensation" technique for the stICA is introduced in the second iteration of the presented algorithm (see Figure 2). In the second iteration (Figure 7), satisfactory object segmentation results are achieved by a compensation approach, a frame object indexing method, and the postprocessing techniques.

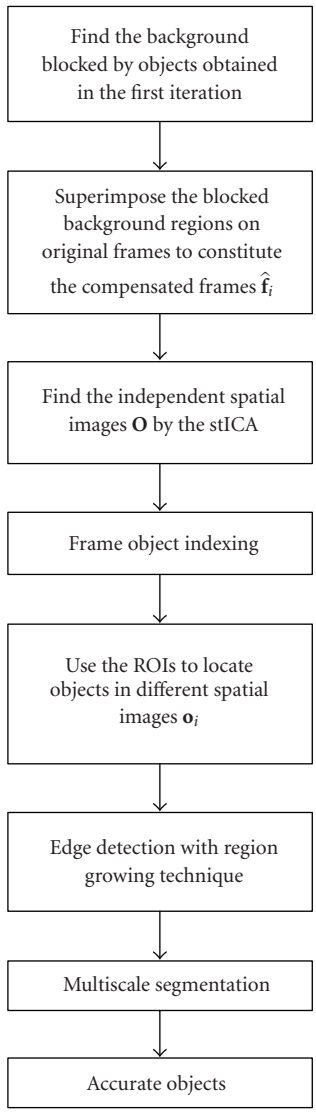


Figure 7: Block diagram of the second iteration.

The second iteration consists of the following procedures (shown in Figure 7):

(1) extracting the regions of background that are blocked by the objects whose boundaries are obtained in the first iteration;
(2) superimposing the regions of background that are blocked by the objects onto the original frames to obtain the compensated frames;
(3) employing the stICA to process the compensated frames to produce spatial signals with clearer edges;
(4) indexing the frame objects by the SVD and the weighting matrices;
(5) using the ROIs obtained in the first iteration to locate the objects in different spatial images;
(6) the postprocessing algorithms, such as edge detection with region growing, and multiscale image segmentation, are applied again to obtain more accurate objects.

Simulation results will illustrate that the proposed approaches along with the postprocessing techniques can segment the objects of interest accurately and effectively.

4.1. A compensation approach of stICA

The major problem of application of the stICA to video sequences is the nonlinear combination problem as shown in (17). The nonlinear problem may lead to the poor outputs from the stICA. Let $\hat{\Delta}_i$ denote the estimation of blocked background region Δ_i in each frame $\mathbf{f}_i$. If we "compensate" the blocked background back to each frame in (17), we can obtain the ideal frames $\hat{\mathbf{f}}_i$ for the linear stICA model:

$$\mathbf{f}_i + \hat{\Delta}_i = \hat{\mathbf{f}}_i + \hat{\Delta}_i - \Delta_i = \hat{\mathbf{f}}_i + (\hat{\Delta}_i - \Delta_i), \tag{27}$$

where Δ_i, $\hat{\Delta}_i$, $\mathbf{f}_i$, and $\hat{\mathbf{f}}_i$ are the $M \times 1$ column vectors as stated in Section 2.

If Δ_i is ideally located, $\hat{\Delta}_i - \Delta_i = \mathbf{0}$, which means that the video frames can fit the stICA model. In fact, if we get the accurate blocked background information, we can outline the objects of interest and fulfil the video object segmentation task. However, we can only acquire the approximate blocked background information in the first iteration and use it for the stICA processing in the second iteration. The following steps are the procedures of the compensated frames for the stICA processing in the second iteration.

(1) The blocked regions of the background are determined by the segmented objects in the first iteration. The blocked regions are used as binary masks that are applied to the background image obtained in the first iteration to estimate the blocked background information $\hat{\Delta}_i$.

(2) The estimated blocked background $\hat{\Delta}_i$ is superimposed onto its corresponding original video frame and the compensated frames are obtained.

Note that here we only deal with the background compensation caused by nonlinear blocking. In general, it is assumed that the objects in the selected frames for stICA do not overlap with each other. We can reasonably achieve this by randomly selecting the raw frames for stICA. If the two moving objects do overlap, the stICA actually treats them as one object and they will be separated together. In such cases, if we want to separate the overlapped individual objects, additional domain knowledge is necessary and the iterative compensation principle may still be used.

4.2. Frame object indexing

Due to the ambiguities of the ICA [26], the order of the ICs after stICA cannot be determined. The order of the ICs is very important for reconstructing the temporal video sequence containing only the segmented objects. Thus, before edge detection, the recovered spatial objects $\mathbf{O}$ must be indexed according to the order of the video frame $\hat{\mathbf{F}}$. In this subsection, an indexing method based on the SVD [36] and the corresponding weighting matrices is developed.

According to (18), the SVD of the video sequence $\hat{\mathbf{F}}$ is

$$\hat{\mathbf{F}} = \mathbf{U}\mathbf{D}\mathbf{V}^T = (\mathbf{U}\mathbf{D}^{1/2})(\mathbf{V}^T\mathbf{D}^{1/2}) = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^T. \quad (28)$$

Since both $\mathbf{U}$ and $\mathbf{V}$ are orthogonal [21], we could make use of these two equations $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{D}^{1/2}$. We then obtain

$$\hat{\mathbf{F}}\mathbf{V} = \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V} = \mathbf{U}\mathbf{D} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \tilde{\mathbf{U}}\mathbf{D}^{1/2}, \quad (29)$$

where $\mathbf{D}$ is a diagonal matrix with singular values. The multiplication of $\tilde{\mathbf{U}}$ with $\mathbf{D}^{1/2}$ can only change the amplitude of $\tilde{\mathbf{U}}$ (eigenimages), but cannot change the eigenimage indices. Let us suppose that $\mathbf{V}$ is a $k \times k$ weight matrix. Eigenimage $\mathbf{u}_i$ $(i = 1, \ldots, k)$ is most affected by the frame $\hat{\mathbf{f}}_i$ that has the largest absolute element in the corresponding column of $\mathbf{V}$, that is,

$$\hat{\mathbf{f}}_i \longleftrightarrow \mathbf{u}_j, \quad j = \arg\max_j \{v_{i,j}, \ \forall i\}. \quad (30)$$

Note that the eigenimages $\mathbf{U}$ and $\tilde{\mathbf{U}}$ have the same image orders. In this way, the relationships between frames and the eigenimages can be obtained.

Referring to (5), the spatial IC images $\mathbf{O}$ (containing objects) can be written as

$$\mathbf{O} = \tilde{\mathbf{U}}\mathbf{W}_O, \quad (31)$$

where $\mathbf{W}_O$ is a $k \times k$ unmixing matrix. The indexing relationship between $\mathbf{U}$ and $\mathbf{O}$ can then be found in the same manner as that used for $\hat{\mathbf{F}}$ and $\mathbf{U}$, that is,

$$\mathbf{o}_i \longleftrightarrow \mathbf{u}_j, \quad j = \arg\max_j \{w_{O,\{i,j\}}, \ \forall i\}. \quad (32)$$

Combining (30) and (32), the indexing relationship between the frames $\hat{\mathbf{F}}$ and the spatial IC images $\mathbf{O}$ can be established. Note that the same object indexing method can be used in both the first and second iterations.

5. SYSTEM SIMULATIONS

In the following illustrations, a grayscale video sequence "Hall Monitor" of 9.28-second duration is used for experiments. There are altogether 280 frames, each with 240×360 pixels and 256 grayscale levels. We suppose that every video frame contains at least one object of interest. This means there is no pure "background" image.

5.1. Simulation of the stICA applied to video processing in the first iteration

A set of frames are selected from the 280 frames for further processing. To avoid interference between close objects, frames are selected from the sequence at a constant interval. We set up a graphic user interface (GUI) that can show the processing details step by step (Figure 8). The program allows users to define a frame selection interval. Based on the frame selection rate, a number of frames are selected from the 280 frames and the stICA model is applied to them. Through the stICA processing, we obtain the same number of spatial output images as input frames. For simplicity of the following illustration, 4 frames are selected as shown in Figure 9.

Among the output images in Figure 10, only the background image (Figure 10(a)) is relatively clear. Meanwhile, other output images (Figures 10(b), 10(c), and 10(d)) contain moving objects but with some undesired shadows. The reason is that the pixels representing objects in the video frames are not the linear combination of the pixels representing objects and the background in recovered image signals. In other words, these video frames are not a linear mixture of all the independent sources, namely the objects and background. Since the background image is relatively clear among all the outputs, it can be subtracted from all original video frames to get the preliminarily processed images which contain only objects as shown in Figure 11.

In these images, we can see extensive noise. Postprocessing techniques are thus required to refine the object segmentation.

5.2. Simulations of the postprocessing techniques in the first iteration

5.2.1. Simulation of wavelet analysis to locate ROIs

After the subtraction of the recovered background, the preliminarily processed images contain object but with extensive noise (e.g., Figures 11(a), 11(b), 11(c), and 11(d)). The DWT decomposes an image into four subspaces: three wavelet subspaces (LH, HL, and HH) and one scaling subspace (LL). A scaling subspace (LL) example is shown in Figure 12(a). It is a low-frequency approximation of the original image. The other three subspaces LH, HL, and HH are shown in Figures 12(b), 12(c), and 12(d). It can be seen that the LH, HL, and HH subspaces describe image details along three directions: vertical, horizontal, and diagonal directions, respectively.

The sliding window filtering described in Section 3.2 is then applied to wavelet subspaces. The empirical constant α in (21) is set at 0.685 since it is proved empirically effective in all test images. The rectangle ROI is shown in Figure 13(a). Figure 13(a) also shows that the locations of the ROIs are very accurate and the object of interest is completely included within the rectangular ROI. Figure 13(b) shows the detected ROI with size 131×57. This reduces the computation complexity for further object extraction.

5.2.2. Simulation of edge detection with region growing

The rectangular ROIs detected by the presented object detection method based on the stICA describe the areas of object of interest, but they do not contain exact boundary information of the detected objects. The Canny edge detection technique is applied to these rectangular ROIs. This operation renders a binary image as shown in Figure 14(b). However, in this binary image of the ROI, not all the detected regions belong to the object of interest. For example, in Figure 14(b), besides the moving human object, there are other regions, such as the door. In the ROIs, the target objects are generally

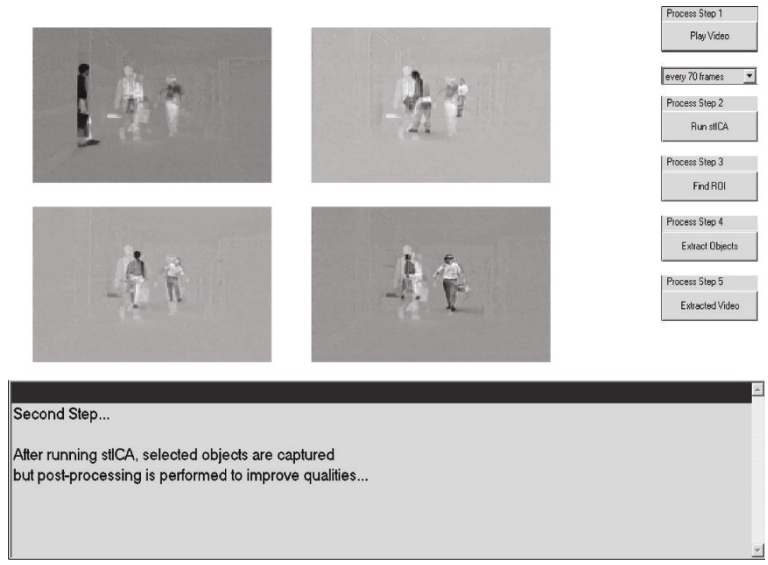


Figure 8: A GUI for the stICA-based object extraction in video sequences.

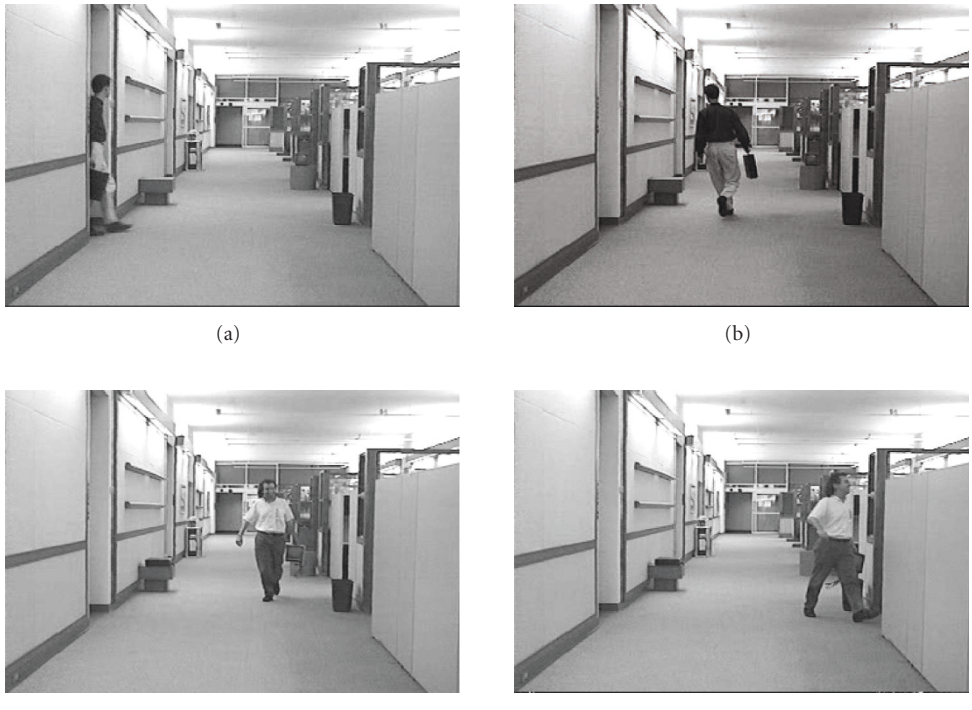
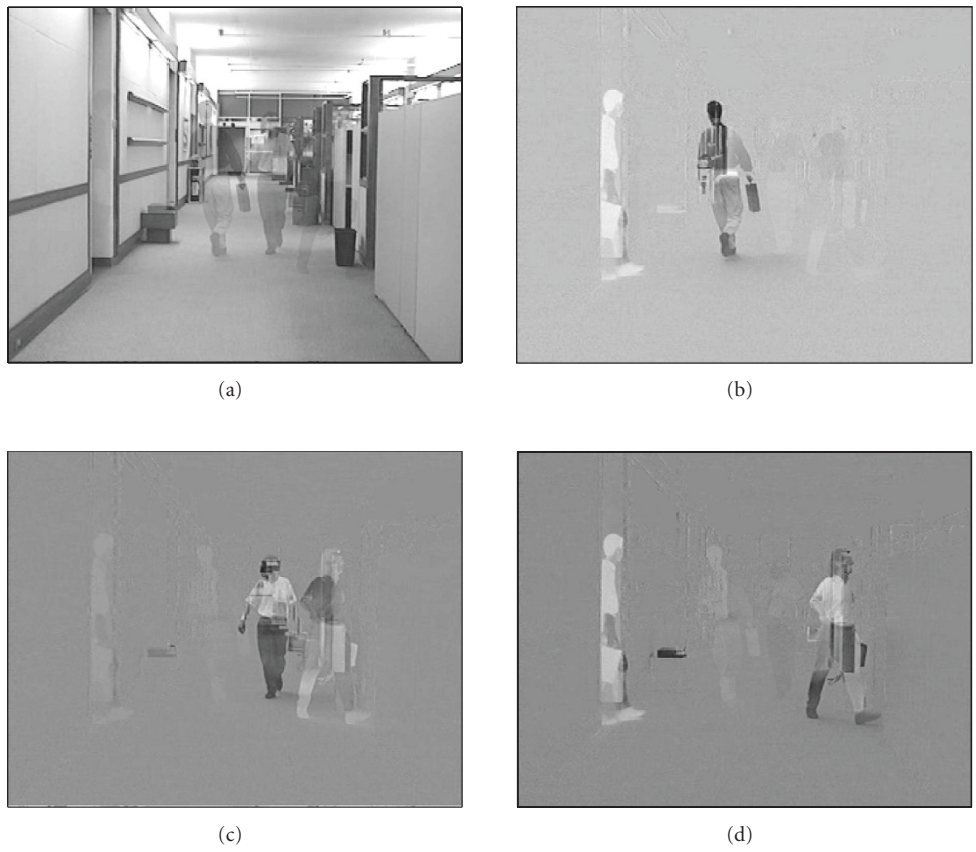


Figure 9: The original video sequence frames.

larger than other isolated regions. Thus we can discriminate the target objects from those unwanted regions through the comparison of their sizes. For example, in Figure 14(b), the size of the moving human is much larger than others.

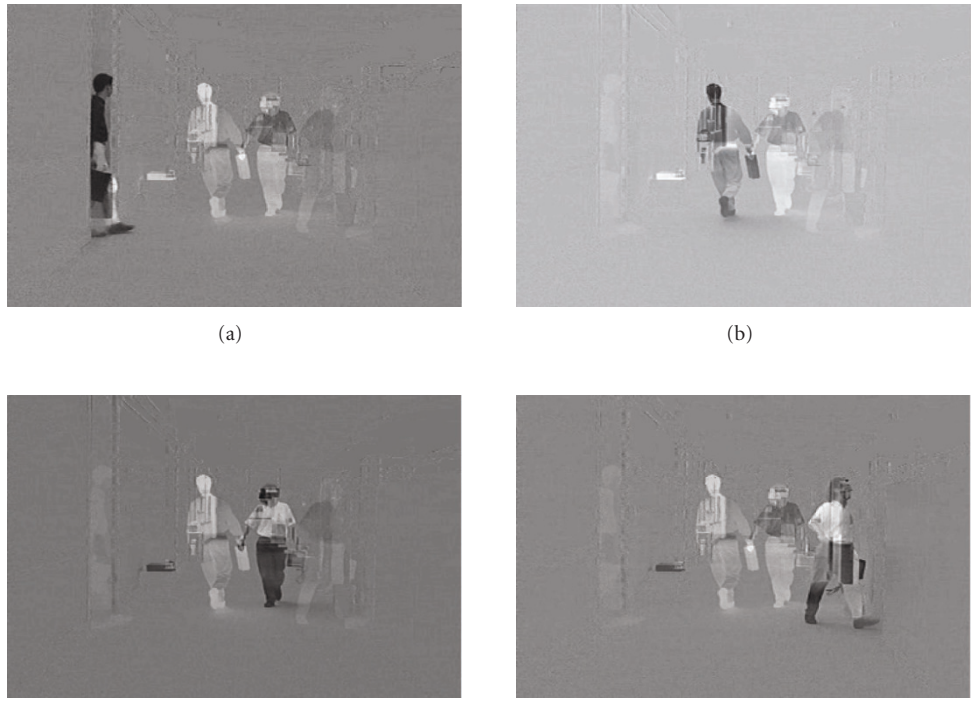
The region growing algorithm described in Section 3.3 is employed to remove the isolated regions that are not the object regions. Figure 14(c) shows three isolated regions. This region growing algorithm is a recursive computing method. Figure 14(d) shows three connected regions that are assigned three labeling integers. The sizes/areas of isolated regions are easily computed. The small regions corresponding to the labeling integers 1 and 3 are eliminated by a region-size threshold detector (Figure 14(e)). This threshold is set to 10% of the largest region size (except the background) in the whole binary image. After threshold detection, only the approximate object of interest remains.



(a) (b)

(c) (d)

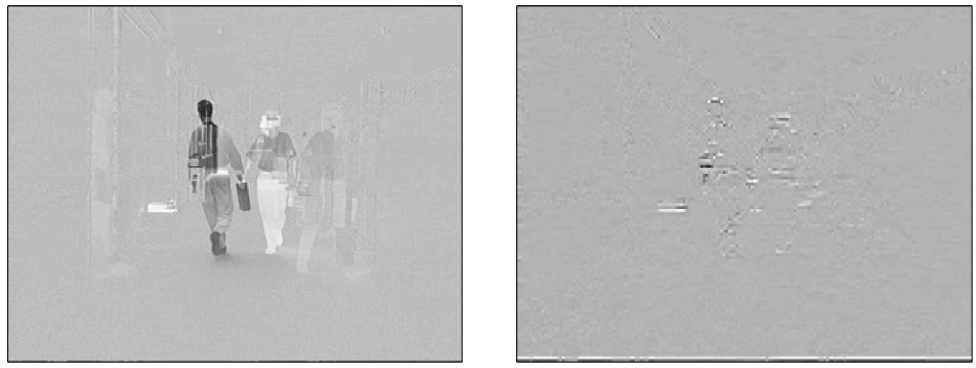
Figure 10: Spatial source signals from the first stICA processing.



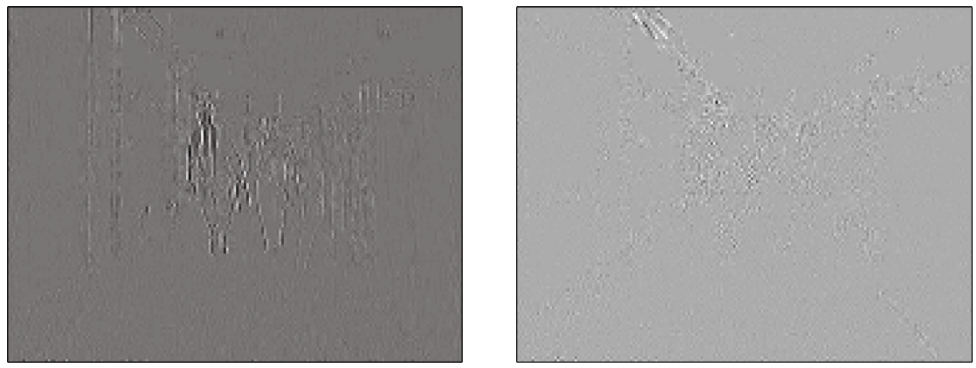
(a) (b)

(c) (d)

Figure 11: Preliminarily processed images from the first stICA processing subtraction.

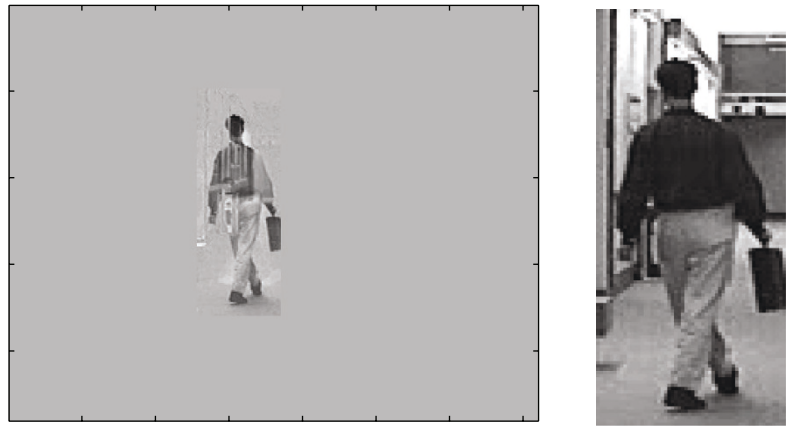


(a) (b)



(c) (d)

Figure 12: An example of 2D wavelet decomposition: (a) LL scaling subspace; (b) LH subspace; (c) HL subspace; (d) HH subspace.



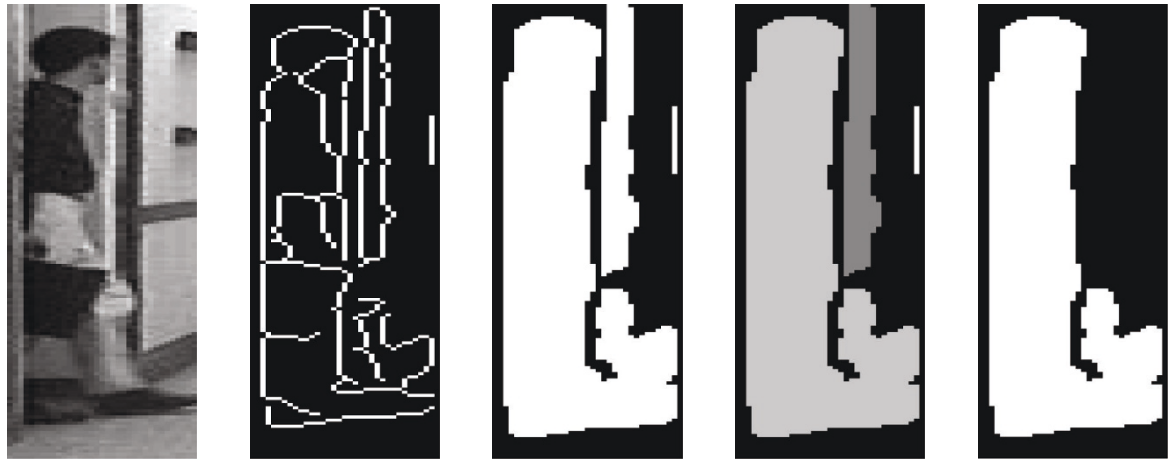
(a) (b)

Figure 13: (a) A rectangular ROI after the horizontal and vertical wavelet analysis; (b) the "zoom-in" video frame.

5.2.3. *Simulation of multiscale image segmentation*

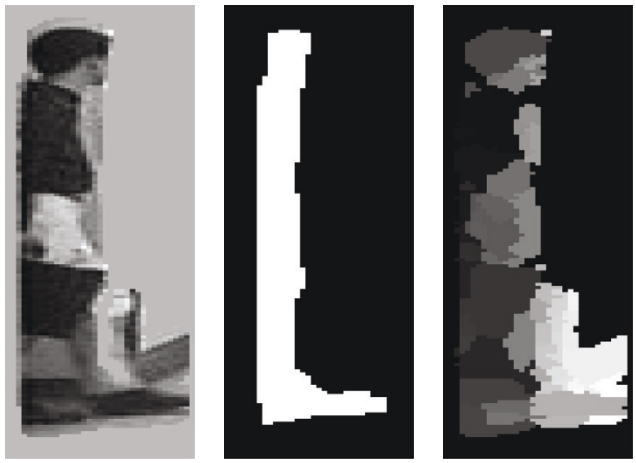
In Figure 14(e), the object regions are obtained by edge detection and region growing. However, this approach cannot remove the superfluous components that are connected to the objects. Such components are caused by the false edges from edge detection.

The multiscale region-based still-image segmentation method outlined in Section 3.4 is employed on the object regions in postprocessing. The superfluous connected components can be removed by eroding [30] the edge-detected regions. After eroding the regions in Figure 14(e), that is, Figure 15(a), a "slimmer" object is obtained and shown in Figure 15(b). We project the pixels after the eroding



(a) (b) (c) (d) (e)

Figure 14: (a) Original image in the ROI; (b) edge detection by the Canny detector; (c) filling regions after edge detection; (d) labeling regions with the same integer; (e) removing regions that are not of interest by threshold detection.

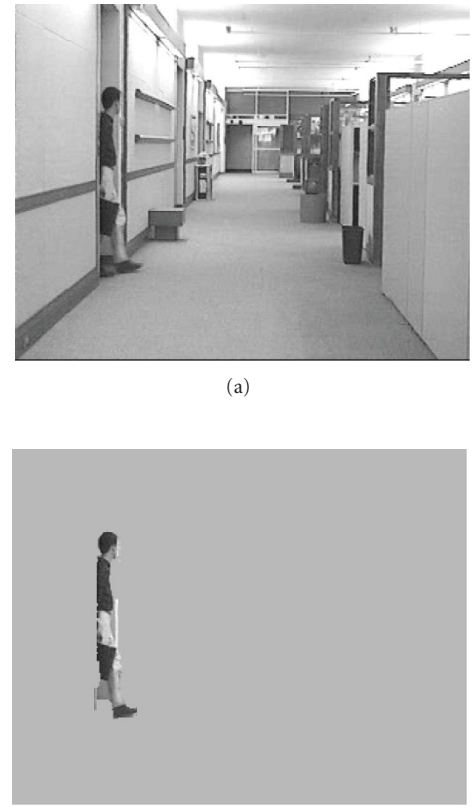


(a) (b) (c)

Figure 15: (a) Object regions in the ROI; (b) eroded regions from edge detection; (c) multiscale segmented regions.

operation to the multiscale segmented image (shown in Figure 15(c)). The regions belonging to the object are identified. The reason for eroding the binary regions is to make sure that no pixel is projected to the connected components. An example of the extracted object from the original image is illustrated in Figure 16(b).

The methods we used in the first iteration work effectively for the objects with a high-contrast clear background, which means that the grayscale of the background pixels is not similar to the target objects. Figures 17(a) and 17(d) are in this category. However, if the background and the object of interest have similar grayscale values, false regions may be identified as the objects of interest, as shown in Figures 17(b) and 17(c), due to the linear assumption in the stICA model.



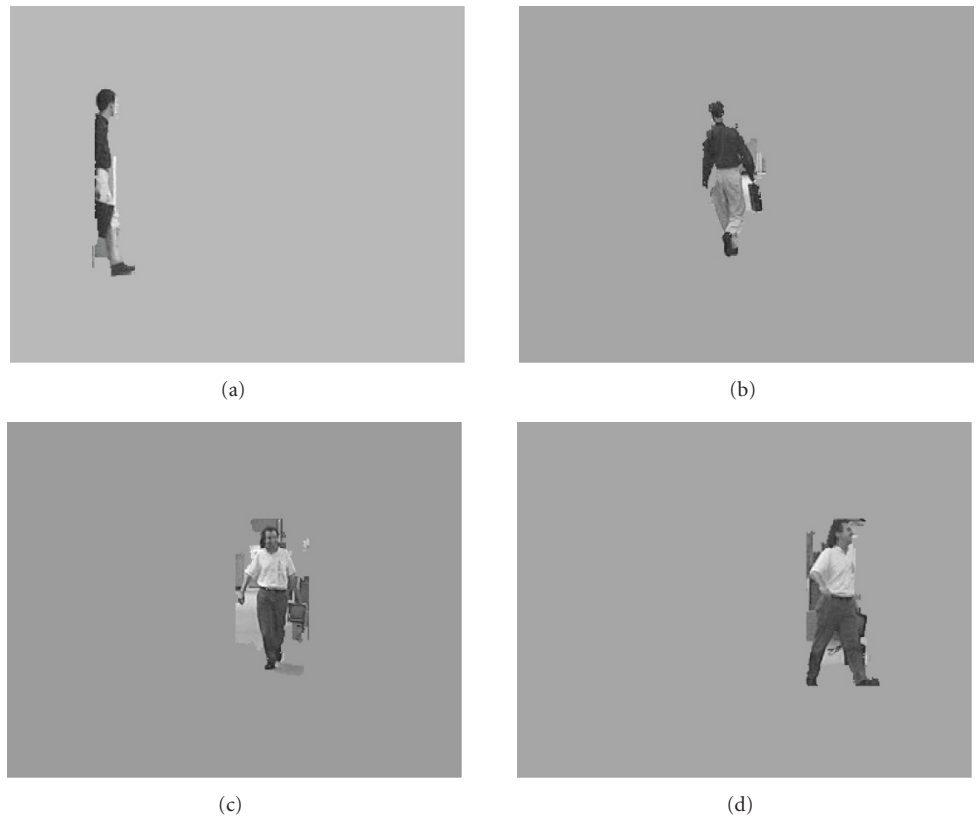
(a)

(b)

Figure 16: (a) Original video frame; (b) extracted object.

5.3. Simulation of compensation approach of stICA

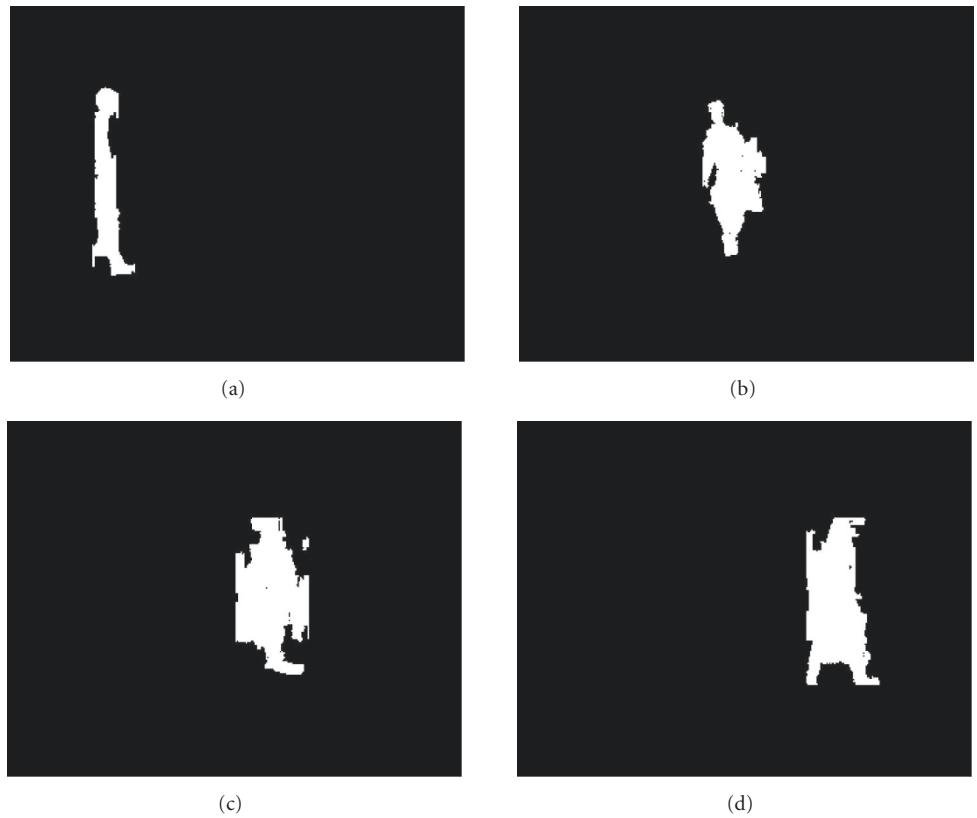
Figures 18(a), 18(b), 18(c), and 18(d) show the binary masks that are determined by the segmented objects from the first iteration. The blocked background regions are obtained by



(a) (b)

(c) (d)

Figure 17: The output images from the first iteration.



(a) (b)

(c) (d)

Figure 18: Binary masks determined by the first iteration.

projecting the masks to the background we recovered in the first iteration as shown in Figure 19. The compensated video frames are the sum of original video frames and the corresponding blocked background regions. Figure 20 provides the examples of the video frames with compensated background. Then the stICA model is applied to the compensated video frames (Figure 20). Figure 21 shows the results of the stICA outputs. As expected, the second stICA processing detects the object edges accurately. Compared with the results obtained in the first stICA processing (Figure 10), the edges of the recovered spatial ICs in the second stICA processing (Figure 21) are clearer and sharper.

5.4. Simulation of the frame object indexing approach

Due to the inherent permutation ambiguity of stICA, the order of the ICs cannot be determined. However, the order of ICs is very important for reconstructing the video sequence containing only the objects.

In the experiment, there are altogether four video frames defined as inputs to the stICA. Since the SVD is the preprocessing tool of the ICA, we first use the SVD to find the indexing relationship between the video frames $\mathbf{F}$ and the eigenimage $\tilde{\mathbf{U}}$. In the eigenimages matrix $\tilde{\mathbf{U}}$, the first principle component $\mathbf{u}_1$ represents the strongest energy among all the principle components [36]. Among all the objects, the background has the strongest energy because it exists in every frame of the video sequence. Thus $\mathbf{u}_1$ should correspond to the background (a special object). Through the observation of the elements of the eigenmatrix $\mathbf{V}$, the indices of other objects can be found.

In the example, there are altogether four objects and one background to be indexed. To determine the indices of the four objects, the largest absolute coefficients in columns two to four of the eigenmatrix $\mathbf{V}$ are found:

$$\mathbf{V} = \begin{bmatrix} 0.4899 & 0.4343 & \underline{-0.7464} & 0.0245 \\ 0.4820 & -0.1482 & 0.4756 & \underline{-0.7302} \\ 0.4948 & \underline{0.6290} & 0.4348 & 0.4129 \\ 0.5019 & 0.2218 & -0.1661 & \underline{-0.8002} \end{bmatrix}. \tag{33}$$

According to (30), the third coefficient of column two has the largest absolute value in that column, indicating that the object segmented from the second eigenimage $\mathbf{u}_2$ will be indexed as the third frame in the video sequence. This is true because the third frame corresponds to the third coefficient and the frame has the largest contribution to the formation of the second eigenimage and to the object in it. For the same reason, the object segmented from the third eigenimage $\mathbf{u}_3$ will be indexed as the first frame in the video sequence. Finally, column four has two large coefficients at positions two and four, indicating that there are two objects to be segmented from the fourth eigenimage $\mathbf{u}_4$ and their indices in the video sequence will be the second and the fourth frames, respectively.

The indexing relationship between the eigenimages $\tilde{\mathbf{U}}$ and the video frames $\mathbf{F}$ can be described as follows:

$$\mathbf{u}_3 \longrightarrow \mathbf{f}_1, \quad \mathbf{u}_4 \longrightarrow \mathbf{f}_2, \quad \mathbf{u}_2 \longrightarrow \mathbf{f}_3, \quad \mathbf{u}_4 \longrightarrow \mathbf{f}_4. \tag{34}$$

Then we use the Bell-Sejnowski algorithm in the stICA to optimize the eigenimages $\tilde{\mathbf{U}}$ and obtain the unmixing matrix $\mathbf{W}_O$ such that $\mathbf{O} = \tilde{\mathbf{U}}\mathbf{W}_O$. In the experiment,

$$\mathbf{W}_O = \begin{bmatrix} \underline{-10.9408} & -0.8998 & 2.1762 & -1.4259 \\ -1.9929 & \underline{-38.6003} & 1.4613 & 2.8184 \\ -0.5246 & 0.2471 & \underline{-40.5752} & 2.8608 \\ -1.0995 & 8.3672 & 6.9683 & \underline{35.0712} \end{bmatrix}. \tag{35}$$

For the same reason outlined above (also see (30)), the relationship between $\tilde{\mathbf{U}}$ and $\mathbf{O}$ is

$$\mathbf{u}_1 \longrightarrow \mathbf{o}_1, \quad \mathbf{u}_2 \longrightarrow \mathbf{o}_2, \quad \mathbf{u}_3 \longrightarrow \mathbf{o}_3, \quad \mathbf{u}_4 \longrightarrow \mathbf{o}_4. \tag{36}$$

Thus, we can map the relationship between $\mathbf{F}$ and $\mathbf{O}$ as follows:

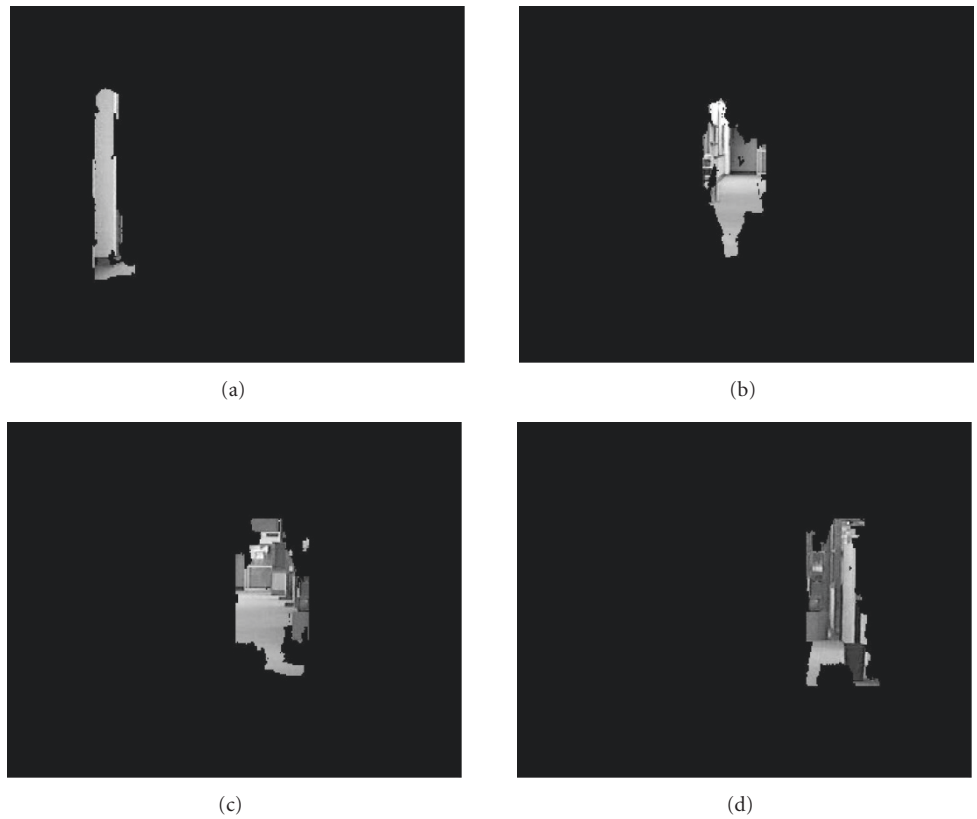
$$\mathbf{o}_3 \longrightarrow \mathbf{f}_1, \quad \mathbf{o}_4 \longrightarrow \mathbf{f}_2, \quad \mathbf{o}_2 \longrightarrow \mathbf{f}_3, \quad \mathbf{o}_4 \longrightarrow \mathbf{f}_4. \tag{37}$$

The object indexing relationship from $\mathbf{F}$ to $\mathbf{O}$ through $\tilde{\mathbf{U}}$ is illustrated in Figure 22. In this way, the frame object order can be determined.

Afterwards, the postprocessing schemes are applied to extract the final object in each frame. The results after the second iteration are shown in Figure 23.

To compare the segmentation image quality in these two iterations, the commonly used peak signal-to-noise ratio (PSNR) [37, 38] is calculated. In the noise calculation, the objects are manually segmented and employed as the true reference segmentation. Table 1 shows the comparison of the PSNR values (dB) of the segmented object images in the two iterations from the "Hall Monitor" sequence. It shows that the results obtained in the second iteration (Figure 23) are superior to those in the first one (Figure 17).

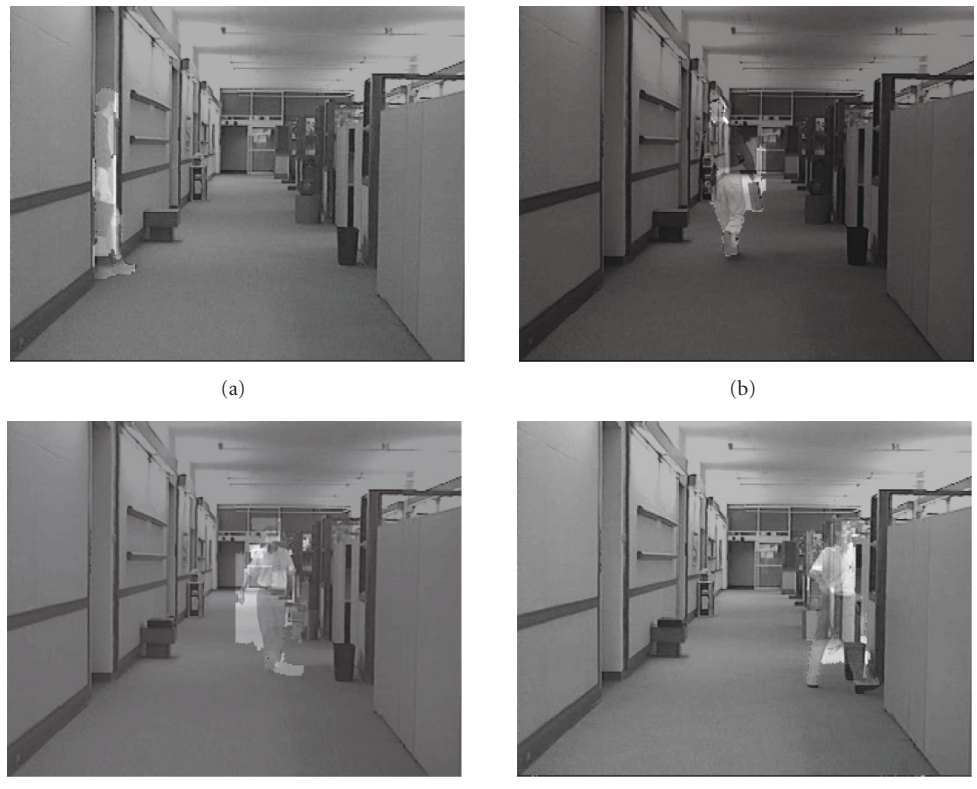
The PSNRs of another simulation experiment are also illustrated below. In this experiment, a "Computer Lab" video sequence of 4.35-second duration is used. There are altogether 160 frames, each of which has 240×360 pixels and 256 grayscale levels. Each video frame contains at least one object of interest, that is, there is no pure "background" image. A set of frames are selected from these 160 frames for processing in the proposed system. At a constant interval of 40, 4 frames are selected to be processed by the system. The two-iteration processing is applied to the selected frames. Table 2 gives the comparison results of the PSNR values of the first and the second iterations. Similar to the results in Table 1, the results after the second iteration are better than the results after the first iteration in Table 2. Moreover, it is found that some missing information of the object in the first iteration can be retrieved back in the second iteration by the proposed compensation method.



(a) (b)

(c) (d)

Figure 19: Blocked background regions determined by the binary mask.



(a) (b)

(c) (d)

Figure 20: Video frames with compensated background for the stICA in the second iteration.

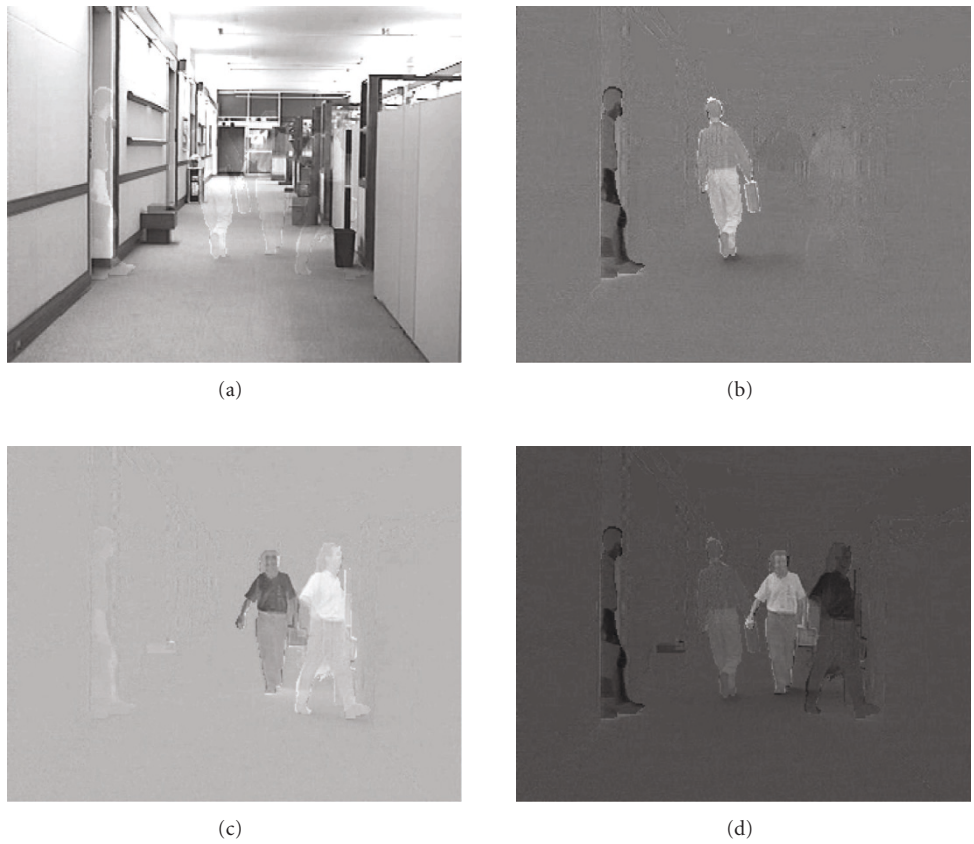


FIGURE 21: Spatial source signals from the second stICA processing: $\mathbf{o}_1$, $\mathbf{o}_2$, $\mathbf{o}_3$, $\mathbf{o}_4$.

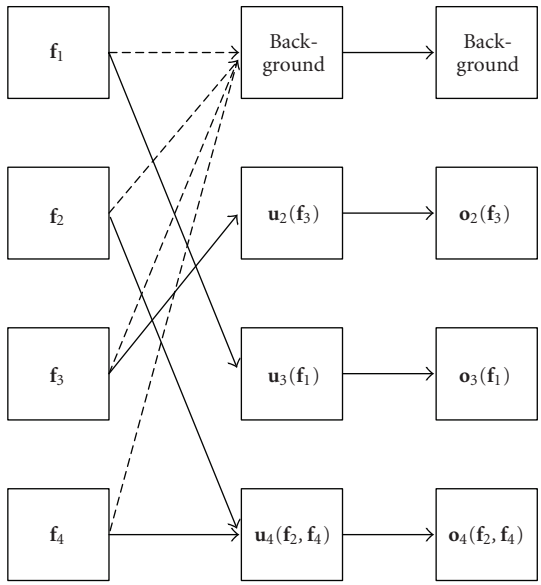


FIGURE 22: Illustration of the indexing relationship from $\mathbf{F}$ to $\mathbf{O}$ through $\tilde{\mathbf{U}}$.

Note that a fixed camera is assumed. Once a background image is extracted, it can be subtracted (by correlation) from all video frames other than the frames used for stICA. Then the postprocessing techniques may be used to remove background noise and variations, and extract the exact objects as well as the relationships of the objects across the frames.

Also note that it is the advantage of the new algorithm that it can still well separate the background even if there is no pure background image since the stICA method can maximally catch the statistical correlation of the background across frames.

5.5. Discussion and practical considerations

The semantic object segmentation has been a challenging topic in video analysis and processing, since there is currently no universal way to define a semantic object using low-level features. This is also a reason that the object-based MPEG-4 coding has not been widely used in applications. The presented new object extraction system is an attempt to employ the joint spatiotemporal statistical features in a video sequence to identify coherent moving objects. We believe that such spatiotemporal features carry reasonable semantic meaning of the objects of interest. However, while the statistical features have advantages to catch large-scale semantic characteristics of objects in video, they are not very accurate identifying details, such as boundaries, of objects. On the other hand, traditional image segmentation methods generally have advantages to catch edges but have difficulties to identify different semantic meaning of various

(a) (b)

(c) (d)

Figure 23: The output images from the second iteration.

Table 1: PSNR (dB) of the segmented images in "Hall Monitor" sequence.

Iteration	Image (a)	Image (b)	Image (c)	Image (d)
First	30.25	27.43	26.12	34.71
Second	36.36	39.84	41.72	40.30

Table 2: PSNR (dB) of the segmented images in "Computer Lab" sequence.

Iteration	Image (a)	Image (b)	Image (c)	Image (d)
First	24.42	29.66	25.17	38.72
Second	26.67	31.21	31.54	40.28

edges. The presented system combines the new spatiotemporal statistical-features-based video analysis method and conventional effective image segmentation methods for video object segmentation. For practical video object segmentation applications, postprocessing steps are generally needed to take advantage of multiple semantic features of videos to obtain accurate segmentation results.

Nevertheless, there are weaknesses for the presented system. First, the current method only deals with static background. Static background provides us with more spatiotemporal information since it leads to more statistical spatiotemporal correlations across frames. Though the static background is assumed, some variations in the background, such as very common background luminance changes and additive noises that happen in a number of applications, can be dealt with well by the stICA model since the stICA method can maximally catch the statistical relationship of the coherent objects across frames. Theoretically, the stICA model has potential inprocessing moving background as well, since the background can be considered as another independent moving object. However, since foreground objects always occlude background, with a moving background and multiple moving objects, more sophisticated algorithms have to be developed to solve the nonlinearity and dependency problems. It is expected that the statistical-modeling-based method can be combined with many other traditional methods to achieve better content analysis of video sequences, a topic that will also be our future investigation. Secondly, many statistical analysis and learning methods, such as ICA methods, can be computationally expensive since nonlinear numerical optimization is usually involved. However, in the presented system, it is not necessary to perform stICA on all frames since backgrounds as well as moving objects among consecutive frames within a video shot or a video scene are highly correlated. After the static background and basic moving objects are identified by stICA in the selected frames, correlation or

other simple algorithms can be developed to identify the objects in the adjacent frames. Also, the stICA algorithm can be further optimized by employing other statistical information. In the present work, our main objective is to demonstrate the validity of the new stICA-based method. The optimization of the algorithms and the development of the real-time analysis are certainly important involved topics that should be further investigated in theory and practice.

In the new system, some empirical parameters, such as the number of frames used for stICA K, have to be selected. In general, K should be no less than the number of moving objects that appeared in the image sequence. However, a large K may lead to increased computational complexity. As the objective of stICA is to identify common background and rough object areas, in practice, an empirical K can be selected as 4 or 5, which gives reasonable results.

6. CONCLUSIONS

In this paper, a new automated video object extraction system is presented based on the stICA and multiscale analysis. A novel statistical formulation based on stICA is proposed for video sequence analysis to extract moving objects. A mathematical framework is presented in the context of the video frame analysis. An advantage of this statistical analysis method is that it captures both the spatial and temporal characteristics of moving video objects in frames without getting into the detailed pixel-based processing. On the other hand, though the new statistical method can catch the moving blobs in the video, it cannot capture the object details in the pixel level. Therefore, a set of postprocessing schemes incorporating traditional pixel-based processing techniques, such as edge detection, region growing, and so forth, are presented to extract the boundary details of objects. Specifically, multiscale analysis is employed in finding the ROIs and segmenting homogenous regions. However, the inherent nonlinearity of the video object composition in a video frame contradicts with the linearity in the ICA model. A new iterative background-compensation scheme is presented to solve this problem.

Extensive experiments are performed to validate the presented model, system, and new algorithms. It is shown that for fixed camera video sequences, the extraction results are satisfactory. It is also worth to note that for the background compensation, although more iterations are possible to further improve the validation of the linear stICA model, one iteration has worked reasonably well in our experiments. Both visual and PSNR results demonstrate the effectiveness of the new system. It is expected that the presented stICA-based object segmentation system, combined with other information processing technologies, can be used in applications such as video information mining, analysis, and retrieval, and so forth.

ACKNOWLEDGMENT

This work was supported by the Canadian Natural Sciences and Engineering Research Council (NSERC) under Grant RGPIN239031 and by Micronet.

REFERENCES

[1] MPEG Video Group, "Mpeg-4 video verification model version 11.0," *ISO/IEC JTC1/SC29/WG11 MPEG98/N2172*, March 1997.

[2] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 625–638, 1994.

[3] G. D. Borshukov, G. Bozdagi, Y. Altunbasak, and A. M. Tekalp, "Motion segmentation by multistage affine classification," *IEEE Transactions on Image Processing*, vol. 6, no. 11, pp. 1591–1594, 1997.

[4] G. Adiv, "Determining three-dimensional motion and structure from optical flow generated by several moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, no. 4, pp. 384–401, 1985.

[5] D. W. Murray and B. F. Buxton, "Scene segmentation from visual motion using global optimisation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 2, pp. 220–228, 1987.

[6] F. Moscheni, S. Bhattacharjee, and M. Kunt, "Spatio-temporal segmentation based on region merging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, pp. 897–915, 1998.

[7] A. Neri, S. Colonnese, G. Russo, and P. Talone, "Automatic moving object and background separation," *Signal Processing*, vol. 66, no. 2, pp. 219–232, 1998.

[8] C. Kim and J.-N. Hwang, "Fast and automatic video object segmentation and tracking for content-based applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, pp. 122–129, 2002.

[9] T. Papadimitriou, K. I. Diamantaras, M. G. Strintzis, and M. Roumeliotis, "Video scene segmentation using spatial contours and 3-D robust motion estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 4, pp. 485–497, 2004.

[10] T. Meier and K. N. Ngan, "Automatic segmentation of moving objects for video object plane generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 525–538, 1998.

[11] T. Meier and K. N. Ngan, "Video segmentation for content-based coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1190–1203, 1999.

[12] H. Xu, A. A. Younis, and M. R. Kabuka, "Automatic moving object extraction for content-based applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 6, pp. 796–812, 2004.

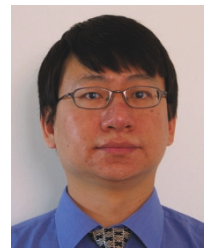
[13] Y.-H. Jan and D. W. Lin, "Extraction of video objects by combined motion and edge analysis," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '02)*, vol. 5, pp. 677–680, Scottsdale, Ariz, USA, May 2002.

[14] J. Pan, S. Li, and Y.-Q. Zhang, "Automatic extraction of moving objects using multiple features and multiple frames," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '00)*, vol. 1, pp. 36–39, Geneva, Switzerland, May 2000.

[15] S. Sun, D. R. Haynor, and Y. Kim, "Semiautomatic video object segmentation using VSnakes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 1, pp. 75–82, 2003.

[16] C. Gu and M.-C. Lee, "Semiautomatic segmentation and tracking of semantic video objects," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 572–584, 1998.

[17] D. Zhong and S.-F. Chang, "An integrated approach for content-based video object segmentation and retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1259–1268, 1999.
[18] D. Gatica-Perez, M.-T. Sun, and C. Gu, "Multiview extensive partition operators for semantic video object extraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 7, pp. 788–801, 2001.
[19] M. J. McKeown, T.-P. Jung, S. Makeig, et al., "Spatially independent activity patterns in functional MRI data during the Stroop color-naming task," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 3, pp. 803–810, 1998.
[20] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
[21] J. V. Stone, J. Porrill, C. Buchel, and K. Friston, "Spatial, temporal, and spatiotemporal independent component analysis of fMRI data," in *Proceedings of 18th Leeds Statistical Research Workshop on Spatial-Temporal Modeling and Its applications*, R. G. Aykroyd, K. V. Mardia, and I. L. Drydent, Eds., pp. 23–28, Leeds, UK, July 1999.
[22] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural networks model," in *Proceedings of International Conference on Neural Networks for Computing*, pp. 206–211, Snowbird, Utah, USA, April 1986.
[23] R. O. Hill, *Elementary Linear Algebra*, Academic Press, Orlando, Fla, USA, 1986.
[24] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.
[25] T. Lee and M. Girolami, "Independent component analysis using an extended informax algorithm for mixed sub-gaussian and super-gaussian sources," in *Proceedings of 4th Annual Joint Symposium on Neural Computation*, vol. 7, pp. 132–139, Los Angeles, Calif, USA, May 1997.
[26] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, New York, NY, USA, 2001.
[27] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, Pa, USA, 1996.
[28] T.-C. Hsung, D. P.-K. Lun, and W.-C. Siu, "Denoising by singularity detection," *IEEE Transactions on Signal Processing*, vol. 47, no. 11, pp. 3139–3144, 1999.
[29] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 617–643, 1992.
[30] V. Hlavac, M. Sonka, and R. Boyle, *Image Processing, Analysis and Machine Vision*, PWS Publishing, Boston, Mass, USA, 2nd edition, 1999.
[31] A. D. Marshall and R. R. Martin, *Computer Vision, Models and Inspection*, World Scientific Publishing, River Edge, NJ, USA, 1993.
[32] M. Tabb and N. Ahuja, "Multiscale image segmentation by integrated edge and region detection," *IEEE Transactions on Image Processing*, vol. 6, no. 5, pp. 642–655, 1997.
[33] X.-P. Zhang, "Target segmentation and extraction from geographic images based on multiscale analysis," in *Proceedings of 5th WSES/IEEE World Multiconference on Circuits, Systems, Communications & Computers (CSCC '01)*, Rethymnon, Greece, July 2001.
[34] X.-P. Zhang, "Multiscale tumor detection and segmentation in mammograms," in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI '02)*, pp. 213–216, Washington, DC, USA, July 2002.
[35] X.-P. Zhang and M. D. Desai, "Segmentation of bright targets using wavelets and adaptive thresholding," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 1020–1030, 2001.
[36] D. C. Lay, *Linear Algebra and Its Applications*, Addison-Wesley, Boston, Mass, USA, 1993.
[37] I.-M. Kim and H.-M. Kim, "A new resource allocation scheme based on a PSNR criterion for wireless video transmission to stationary receivers over Gaussian channels," *IEEE Transactions on Wireless Communications*, vol. 1, no. 3, pp. 393–401, 2002.
[38] S. Saha and R. Vemuri, "An analysis on the effect of image features on lossy coding performance," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 104–107, 2000.



Xiao-Ping Zhang received the B.S. and Ph.D. degrees from Tsinghua University, in 1992 and 1996, respectively, all in electronic engineering. Since Fall 2000, he has been with the Department of Electrical and Computer Engineering, Ryerson University, where he is now an Associate Professor and Director of Communication and Signal Processing Applications Laboratory (CASPAL). Prior to joining Ryerson, from 1996 to 1998, he was a Postdoctoral Fellow at the University of Texas, San Antonio and then at the Beckman Institute, the University of Illinois at Urbana-Champaign. He held research and teaching positions at the Communication Research Laboratory, McMaster University, in 1999. From 1999 to 2000, he was a Senior DSP Engineer at SAM Technology, Inc., at San Francisco, and a Consultant at San Francisco Brain Research Institute. His research interests include signal processing for communications, multimedia data hiding, retrieval, and analysis, computational intelligence, and various applications in bioengineering and bioinformatics. He is the Publicity Cochair for ICME '06 and Program Cochair for ICIC '05. He received Science and Technology Progress Award by State Education Commission of China, for his significant contribution in a National High-Tech Project in 1994. He is a registered Professional Engineer in Ontario, Canada, and a Senior Member of the IEEE.

Zhenhe Chen received the B.E. degree in electrical engineering from South China University of Technology, Guangzhou, China, in 1996 and the M.A.Sc. degree in electrical and computer engineering from Ryerson University, Toronto, Canada, in 2003. He is currently pursuing the Ph.D. degree in electrical and computer engineering at the University of Western Ontario. From 1996 to 2000, he worked for State China Administration of Taxation as a Project Coordinator of Golden Tax of Chinese Government. His research interests are video segmentation by independent component analysis, multiple view geometry, and robot navigation within probabilistic frameworks. He has been a reviewer for conferences in his area of research.